ABSTRACT

Title of Dissertation: MARKET STRUCTURE AND CONGESTION EXTERNALITIES: THEORY AND APPLICATION TO THE RIDE-HAILING INDUSTRY

Julian Andres Gomez Gelvez
Doctor of Philosophy, 2021

Dissertation Directed by: Professor Roberton C. Williams III
Department of Agricultural and
Resource Economics

The encompassing theme of this dissertation is the analysis of markets that feature market power and negative externalities. I focus mainly on congestion externalities, as externalities that affect only market participants.

The first chapter evaluates the efficiency of private pricing of congestible resources. I develop a model of congestible resource use that explicitly considers a bivariate distribution of reservation values and sensitivities to congestion across potential users. This model highlights the importance of the correlation between reservation values and sensitivities to congestion to judge the efficiency of private pricing. Numerical results based on a road pricing example show that monopolistic pricing can range from very inefficient (price too high) when the correlation is negative to almost complete

efficiency when it is strong and positive.

The second chapter studies ride-hailing markets mediated by digital platforms like Uber. I extend the model of the first chapter to include a supply side of drivers. A monopolistic platform chooses prices on both sides of the market to maximize profit. I calibrate the model to the morning peak period of Bogotá, Colombia. The results show that the price gap imposed by a monopolistic platform corresponds to about two thirds of the net marginal external cost caused by an additional ride hailer. A congestion charge on ride hailing is then justified. However, the optimal congestion charge, as a tax on the price charged to riders, covers only 50% of the marginal external cost.

The last chapter explores the effects of modifying several assumptions of the ride-hailing model developed in the second chapter. The main modification is to move from a monopolistic market structure to a duopoly. I show that absent any differentiation between platforms, competition leads to zero profits. This result supports the idea that ride-hailing markets gravitate towards a single platform. Assuming a small amount of differentiation, the duopoly equilibrium reduces the price charged to riders and increases the size of the market. This expansion reduces overall welfare due to the external effect on traffic congestion and calls for a higher congestion charge.

MARKET STRUCTURE AND CONGESTION EXTERNALITIES: THEORY AND APPLICATION TO THE RIDE-HAILING INDUSTRY

by

Julian Andres Gomez Gelvez

Dissertation submitted to the Faculty of the Graduate School of the University of Maryland, College Park in partial fulfillment of the requirements for the degree of Doctor of Philosophy

2021

Advisory Committee:

Professor Roberton C. Williams III, Chair/Advisor

Professor James Archsmith

Professor Joshua Linn

Professor Stephen W. Salant

Professor Andrew Sweeting

PREFACE

This preface tells the (usually untold) story of how I arrived at the research ideas contained in this dissertation. Hopefully, this brief story will be useful for current and future doctorate students struggling to come up with their own research ideas.

I started thinking about the topic of this dissertation during the prospectus development class taught by Prof. Stephen Salant in the fall semester of 2017. Prof. Salant proposed an assignment in which students had to come up with a potential research idea out of a newspaper article. I used an article from The New York Times that argued against the usual practice of reserving the left lane of escalators for people walking, proposing instead that people should stand on both lanes.[1] I felt that the studies in which the article was based missed the fact that people walking on escalators are more likely to be in a hurry (in technical terms, have a higher value of time) relative to people standing. Even though forcing everyone to stand decreases the average time people take to go through the escalator, this result is not enough to conclude it is the best policy, because the improvement in average masks the negative impact on hurried walkers. I developed a simple model to explain my point, and this process made me understand the importance of accounting for heterogeneity in values of time in the analysis of congestible resources such as escalators.

At that point, I had recently read a paper coauthored by Prof. Salant which compared social and private incentives to toll congested roads (Salant & Seegert, 2018). Since that paper implicitly assumes a unique value of time across travelers, I naturally explored how their conclusions would change under heterogeneous values of time. Chapter 2 presents the final result of this exploration. I presented an early version of this chapter in 2018 at the World Congress of Environmental and Resource

[1] www.nytimes.com/2017/04/04/us/escalators-standing-or-walking.html

Economics (WCERE) and the Institutional and Organizational Economics Academy (IOEA).

The ideas in Chapter 2 can be applied to different types of congestible resources, such as roads, fisheries and the internet. I briefly explored their application to the internet in the context of net neutrality regulations (should internet service providers be allowed to prioritize time-sensitive content?) in a term paper for an Industrial Organization class. However, my interest and background in transportation led me to focus on the nascent ride-hailing industry mediated by digital platforms like Uber. By reading economic journals, as well as general interest news and blogs, I noted that this industry is prone to suffer from market power and negative externalities, the two ingredients required to apply the theoretical framework of Chapter 2. The application to ride hailing led to Chapter 3. I presented an early version of this chapter in 2019 at the Summer Conference of the Association of Environmental and Resource Economists (AERE) and the Annual Conference of the International Transportation Economics Association (ITEA).

During my prospectus defense in August 2019, I was asked by one of the committee members about the impact on competition between ride-hailing platforms of the fact that riders can easily check on their smartphones the location of the closest idle vehicle from each platform. I did not have an answer at that point, but it was a very interesting question. I worked on this question after the prospectus defense, and this work led to Chapter 4.

There is plenty of good advice on how to come up with and develop research ideas in economics, but my experience leads me to emphasize two simple recommendations. The first one is to choose a topic that you find genuinely interesting. It takes a significant amount of motivation to work on a single topic for several years, and I found that one of my biggest sources of motivation was pure curiosity. The second is to write or present your work regularly, starting from early stages. I can count

over twenty instances in which I prepared an abstract, a paper or a presentation for a meeting with an advisor, a workshop or a conference, and every time this process allowed me to clarify my ideas, identify gaps in my reasoning or interpret my results from a different perspective.

DEDICATION

I dedicate this dissertation to my grandparents. Thanks to their effort, I was able to obtain the education they dreamed of.

ACKNOWLEDGEMENTS

I would like to thank first the members of my dissertation committee, Professors Rob Williams, Josh Linn, Stephen Salant, James Archsmith and Andrew Sweeting, for their guidance, comments and suggestions at different stages. I am especially thankful to my advisor Rob Williams for his additional support with numerous administrative tasks and for his willingness to discuss topics beyond those directly related to my dissertation. I would also like to thank Professors Robert Chambers and Anna Alberini for their guidance and encouragement during the first years of my doctorate program.

I received financial support from Fulbright Colombia and the Colombian Ministry of Science, Technology and Innovation (Minciencias) during the first four years of my doctorate studies, for which I am grateful. I also received important support from the Department of Agricultural and Resource Economics in the form of graduate assistantships for five years, and excellent administrative support from Pin Wuan Lin and Dany Burns.

I made several good friends during my time at Maryland, who made my doctorate experience inspiring and enjoyable. Special thanks to Tina, Jeff, Dave, Tianqi, Yujie, Han, Mehrab, Rosa, José, Aldo, Camila and Gonzalo for all the time we spent together.

Finally, thanks to my family for their constant encouragement during these years and to my wife for her willingness to move with me to the US and her unconditional support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This dissertation is composed of three main chapters. The encompassing theme for all three chapters is the analysis of markets that feature market power and negative externalities. Each of these two distortions by itself causes markets to fail, but they do so in opposite ways. Firms with market power tend to set prices above efficient levels, while negative externalities imply that efficient prices are above competitive levels. As a result, the simultaneous presence of these two distortions creates a challenge for regulation, because it is not clear *a priori* which way the market is distorted. I focus mainly on congestion externalities, as externalities that affect only market participants. Congestion externalities bring an additional element into the picture, as firms with market power partially internalize them.

The main objective of the first chapter (Chapter 2) is to evaluate the efficiency of private pricing of congestible resources. This evaluation is relevant in the design of regulation for markets that feature congestion and market power, such as electricity markets and the internet, and for the analysis of market power as a toll to control over-exploitation of traditional common-pool resources such as oil and fisheries. I develop a model of congestible resource use that explicitly considers a bivariate distribution of reservation values and sensitivities to congestion across potential users. This model highlights the importance of the correlation between reservation values and sensitivities to congestion to judge the efficiency of private pricing. Numerical re-

sults based on a road pricing example show that monopolistic pricing can range from very inefficient (price too high) when the correlation is negative to almost complete efficiency when it is strong and positive. To the extent that income influences reservation values and sensitivities to congestion, their correlation will likely be positive in most applications. As a result, studies that assume no correlation could overestimate the inefficiency of private pricing. Chapter 2 also analyzes a duopoly setting in which users self select into the congestible alternatives based on their sensitivities to congestion. Even though positive correlation also improves the efficiency reached by a duopoly, the effect is less significant because a duopoly always achieves relatively high efficiency.

Chapter 3 turns to the analysis of ride-hailing markets mediated by digital platforms like Uber. These markets feature market power and congestion externalities. On the side of market power, usually one or very few platforms control the market in each city, which allows them to impose profit-maximizing gaps between the prices charged to riders and paid to drivers (i.e. platform commission). On the side of congestion, ride-hailing vehicles exacerbate traffic congestion, increasing travel times not only for ride hailers but also for other road users.

In order to model ride-hailing markets, I extend the model of Chapter 2 to include a supply side of drivers, who decide to enter the market based on expected revenues. A monopolistic platform then chooses prices on both sides of the market with the objective of maximizing profit. I calibrate the model to the morning peak period of Bogotá, Colombia, one of the most congested cities in the world. The results show that the price gap imposed by a monopolistic platform corresponds to about two thirds of the net marginal external cost caused by an additional ride hailer. A congestion charge on ride hailing is then justified. However, the optimal congestion charge, as a tax on the price charged to riders, covers only 50% of the marginal external cost. This optimal charge takes into account the incomplete pass-through of the tax

that results from the monopolistic structure of the market. Even though optimal regulation of ride-hailing markets involves regulation on both sides of the market (riders and drivers), the optimal congestion charge on the side of riders achieves almost all of the available welfare gains, which leaves little motivation for additional regulation on the side of drivers.

The last chapter (Chapter 4) explores the effects of modifying several assumptions of the model in Chapter 3. First, I move from a monopolistic market structure to a duopoly. I show that absent any differentiation between platforms, competition leads to zero profits. This result supports the idea that ride-hailing markets tend to gravitate towards a single platform. Assuming a small amount of differentiation, the duopoly equilibrium reduces the price charged to riders an increases the size of the market. This expansion reduces overall welfare due to the external effect on traffic congestion, and calls for a higher congestion charge. Second, motivated by the advent of autonomous vehicles, I assume platforms choose directly the number of vehicles available for service (i.e. choosing quantities instead of prices on the supply side of the market). This modification lessens the strength of competition between platforms and reduces the size of the market, which increases overall welfare and calls for a smaller congestion charge. Finally, I explore the effect of allowing travelers to choose between platforms based on realized instead of expected wait times. This adjustment reflects the fact that modern ride-hailing platforms allow travelers to check the location of the closest idle vehicle before deciding to hail a ride. Even though this technological feature does not affect the strength of competition between platforms, it does improve the efficiency of rider-driver matching.

Taken together, these three chapters contribute to the economic literatures on environmental regulation under market power (Buchanan, 1969; Fowlie, Reguant, & Ryan, 2016), optimal and private pricing of congestible resources (Mills, 1981; Verhoef & Small, 2004; Salant & Seegert, 2018), and ride-hailing markets (Arnott,

1996; Cramer & Krueger, 2016; Frechette, Lizzeri, & Salz, 2019).

I relegate lengthy analytical derivations to several appendices.

Chapter 2

On the Efficiency of Private Pricing of Congestible Resources

2.1 Introduction

The growing penetration and use of the internet has lead to a modern congestion problem. During peak-load hours, congestion causes delays in the delivery of content (e.g. video streaming or web browsing) by internet service providers (ISPs). ISPs then have an incentive to prioritize time sensitive content such as video streaming in an attempt to obtain higher revenues from price-discriminating both content providers (e.g. Netflix and Zoom) and end users (Greenstein, Peitz, & Valletti, 2016). Even though ISPs argue that prioritization leads to more efficient management of their congested infrastructure, many countries have issued net neutrality regulations banning this practice, partially in fear of how ISPs could exercise their market power to extract higher rents.

The congestion of telecommunications infrastructure is a modern problem. However, the debate about the efficiency of private ownership and pricing of a congestible resource by firms with market power has been around for over a century. In the first edition of his book *The Economics of Welfare*, Arthur C. Pigou (1920) argued for government intervention to reduce the amount of resources dedicated to industries with increasing cost. Pigou illustrated his argument with the example of travelers choosing between a quick but narrow (hence congestible) road and a slow and ample one. If travelers were free to choose between the two roads, too many would choose the

narrow one because their individual decisions do not weight the cost imposed on other travelers by increasing congestion. Pigou recommended a tax on the narrow road, which could be adjusted to achieve an efficient allocation of travelers between the two roads. In response, Frank H. Knight (1924) criticized Pigou's assumption about free access to the congestible road. In a private ownership economy the congestible road would be managed by a private firm, and this firm, Knight reasoned, would charge an access fee equivalent to Pigou's recommended tax, achieving efficiency without the need for government intervention. Pigou deleted the roads example from subsequent editions of his book, presumably to avoid Knight's criticism (Cheung, 1973, footnote 2).[1]

Mills (1981) clarified this debate using a mathematical model of congestible resource use. Mills identified two main distortions between the price set by a private owner and the efficient one. First, the usual market power distortion pushes private prices above efficient levels. The second distortion is more subtle.[2] Given heterogeneity across potential users in their sensitivities to congestion, the private owner internalizes congestion according to its valuation by *marginal* users. Unless marginal users are representative of all users in terms of their sensitivity to congestion, the private owner does not internalize congestion correctly.[3] Mills noted that the direction of this distortion depends on whether users with high reservation values tend to have high or low sensitivities to congestion. If these two dimensions of heterogeneity are negatively correlated, the private owner would over-internalize congestion because the average sensitivity to congestion of marginal users would be higher than that of all users. This additional distortion would then add to market power in pushing pri-

[1] Button (2020) clarifies that Pigou may have deleted the roads example due to other criticisms, including the realism of increasing cost assumptions for industries and the excessive length of the book. It is not clear whether Pigou was aware of Knight's criticism at the time he wrote the second edition of his book.

[2] Buchanan (1956) identified the market power distortion before Mills, but failed to recognize the second distortion.

[3] This type of distortion was first identified by Spence (1975) in his analysis of quality provision by a monopolist.

vate prices above optimal levels. Conversely, if the correlation is positive, the new distortion would attenuate market power and private pricing would be more efficient.

This chapter makes two contributions to the economic literature that evaluates the efficiency of private pricing of congestible resources. First, I develop a static model of congestible resource use that explicitly considers a bivariate distribution of reservation values and sensitivities to congestion in the population of potential users. This model leads to simple analytical conditions for the welfare- and revenue-maximizing access fees, which highlight the two distortions studied by Mills and the importance of correlation for the efficiency of private pricing. Additionally, the model extends easily to two or more congestible alternatives, where users self select into the alternatives according to their sensitivity to congestion. This extension allows for the study of a duopoly setting with endogenous congestion, no inherent differentiation between the two alternatives and Bertrand competition in prices. This duopoly setting is novel. The usual duopoly setting with endogenous congestion employed in the literature is based on inherent differentiation between the two congestible alternatives (see for instance Basso and Zhang (2007) and Silva and Verhoef (2013)).

Second, I explore quantitatively the impact of correlation between reservation values and sensitivities to congestion on the efficiency of private pricing using a numerical example borrowed from the road pricing literature (Verhoef & Small, 2004). The results show that the efficiency of revenue-maximizing pricing by a monopolist varies considerably with the degree of correlation. Private pricing is several times worse than free access under negative correlation, while it can achieve complete efficiency under strong positive correlations. Even though positive correlations also improve the efficiency of a duopoly, this effect is significantly more tenuous. A duopoly achieves high efficiency even in the presence of strong negative correlations.

This quantitative exploration is important because most studies that evaluate the efficiency of private pricing either implicitly or explicitly assume a value for this cor-

relation, which may bias their assessment of private pricing. Most of these studies belong to the transportation literature, where congestion and market power arise frequently. For example, Boffa, Fedele, and Iozzi (2020) study the welfare effects of transitioning from a regime with atomistic drivers to one where all travelers are supplied by a fleet of autonomous vehicles controlled by a monopolist. Their main results are based on the assumption of a positive linear relationship between reservation values and values of time, which probably creates an overly optimistic view of monopolistic pricing. On the other hand, Verhoef and Small (2004) provide an exhaustive analysis of the relative efficiency of private pricing of congestible roads, from which they conclude that private pricing is usually very inefficient. Their results, however, are based on a bivariate distribution of reservation values and values of time that contains almost no correlation. To the extent that there exists a positive correlation, as may be expected from the influence of income levels on both characteristics, their analysis may be overly pessimistic for private pricing. Similarly, several studies assume a unique value of time across travelers (see for instance Brueckner (2002) and Salant and Seegert (2018)), which negates the potentially beneficial effect for private pricing of the second distortion identified by Mills. The current literature on congestible resources lacks empirical studies that identify the correlation between reservation values and sensitivities to congestion.[4]

The domain of congestible resources, however, extends well beyond transportation-related applications. As the introductory example showed, the current debate over net neutrality regulations for the internet hinges on the efficiency of congestion management by firms with market power. Other modern industries that feature congestion and market power include the radio frequency spectrum and the electricity grid (Borenstein, Bushnell, & Stoft, 2000). Traditional common-pool resources such as

[4]I estimate this correlation for peak-hour travelers in Bogotá, Colombia, in Chapter 3. As expected, I find a positive correlation ($\rho = 0.63$) between values of time and reservation values for ride hailing.

oil and fisheries also feature congestion, as an increase in effort by a market participant reduces the productivity of other participants. Although market power could also be used to control the over-exploitation incentives in this type of industries, the economic literature has also explored other approaches. For example, Heintzelman, Salant, and Schott (2009) explore how the formation of revenue-sharing partnerships creates free-riding incentives that can contain over-exploitation.

The rest of this chapter is organized in three sections. Section 2.2 introduces the theoretical model. Section 2.3 presents the numerical results. The last section summarizes the main contributions of this paper to the literature on congestible resources.

2.2 A model of congestible resource use

Consider a population of N potential users of a congestible resource. Each user i derives a value V_i from use of the resource when it is uncongested (reservation value in monetary terms). The value of the outside option for each potential user is normalized at zero. The level of congestion of the resource is a non-decreasing function of the number of users, denoted $g(x)$. Congestion reduces the reservation value of each potential user linearly, but potential users differ in their sensitivity to congestion (i.e. in how fast congestion reduces their reservation value). The sensitivity to congestion of potential user i is β_i. The value that user i derives from use of the resource when there are x total users is then $V_i - \beta_i \cdot g(x)$. If an access fee τ is charged to use the resource, a potential user would use it only if $V_i - \beta_i \cdot g(x) - \tau \geq 0$.

Two parameters, V_i and β_i, identify each potential user. A bivariate density function $f(\beta, V)$ describes the distribution of these two parameters across the population of potential users. The support of this density function lies entirely in the first quadrant (positive reservation values and sensitivities to congestion). Together with the congestion function $g(x)$ and the number of potential users N, the bivariate density

function $f(\beta, V)$ defines any particular specification of the model.

In the case of roads, the level of congestion refers to travel time, while sensitivities to congestion refer to values of time. V_i represents the value a traveler gains from using the road instead of her outside option, which may be not to travel, to use public transit or any other mode of transportation.[5] V_i also takes into account any operational costs for the traveler such as gasoline consumption.

2.2.1 Use under free access

If there is no access fee, how many people would use the resource and who would these users be? Assuming that potential users decide simultaneously whether to use the resource, Nash Equilibrium (NE) is a suitable solution concept for this question. Let us first characterize the users at a given congestion level. At congestion level g, only those potential users for whom $V_i - \beta_i \cdot g \geq 0$ would like to use the resource. These users can be identified in a two-dimensional graph with β in the horizontal axis and V in the vertical one as those above and to the left of a ray passing through the origin with slope equal to the congestion level g. Figure 2.1 illustrates this characterization.

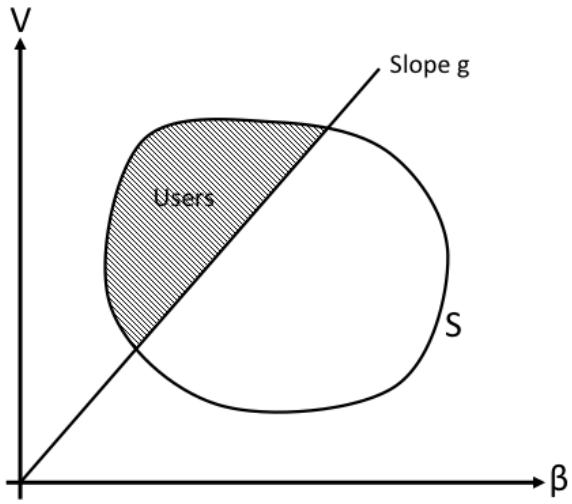
For a congestion level to identify a NE, the number of potential users willing to use the resource at this congestion level must be the same as the number of users required to generate such congestion according to the congestion function $g(x)$. Denoting the number of users in equilibrium under free access as x^0, we can express the previous condition mathematically as

$$N \int_0^{\infty} \int_{\beta g(x^0)}^{\infty} f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta = x^0. \tag{2.1}$$

As long as the congestion function $g(x)$ and the bivariate density function $f(\beta, V)$ are well behaved in terms of continuity and smoothness, a solution to the previous

[5] If the outside options of potential users include other congestible alternatives, I assume they are optimally priced to avoid second-best considerations in the optimal pricing of the congestible resource under study.

Figure 2.1: Users under free access at congestion level g.



Note: S represents the support of the bivariate distribution of sensitivities to congestion (β) and reservation values (V) across potential users.

fixed-point problem exists and is unique. In general, individuals with high reservation values and low sensitivities to congestion will be the ones using the resource in equilibrium.

2.2.2 Use under an access fee

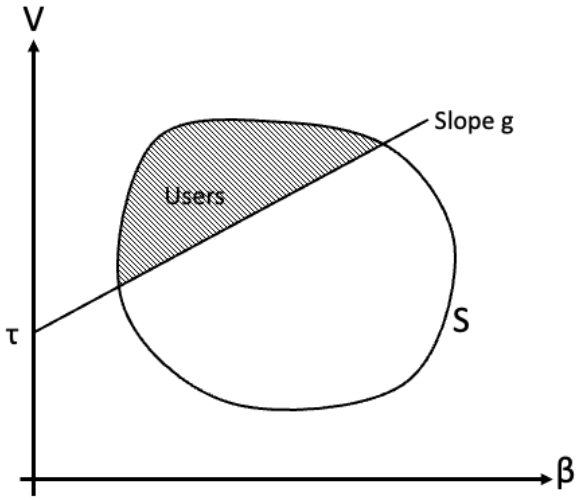
Once a social planner or a private firm imposes an access fee τ, only those potential users for whom $V_i - \beta_i \cdot g - \tau \geq 0$ would like to use the resource at congestion level g. These users can again be identified in a two-dimensional graph as in Figure 2.2.

The condition for a NE changes only slightly in this case to take into account the value of the access fee. Denoting the number of users in equilibrium under an access fee τ as x^τ, the condition is

$$N \int_0^\infty \int_{\tau+\beta g(x^\tau)}^\infty f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta = x^\tau. \tag{2.2}$$

It is not hard to check that $x^0 \geq x^\tau$ and $g(x^0) \geq g(x^\tau)$ for any $\tau > 0$. Assume $g(x^\tau) > g(x^0)$. Then the number of potential users willing to use the resource under access fee τ and congestion level $g(x^\tau)$ (left-hand side of Equation 2.2) would clearly

Figure 2.2: Users under access fee τ at congestion level g.



be less than those willing to use it under free access and the lower congestion level $g(x^0)$ (left-hand side of Equation 2.1). Given the equilibrium conditions, this result implies $x^\tau < x^0$, which is a contradiction with the initial assumption and the fact that the congestion function is non-decreasing.

2.2.3 Optimal allocation of users and access fee

The optimal or efficient allocation of users to the congestible resource maximizes the total economic value created by the resource, which is the sum of the value gained by all its users. Let us start by noting that the equilibrium allocation of users to the resource under free access is not efficient. If one of the marginal users (those for whom $V_i - \beta_i g(x^0) = 0$) stops using the resource, her welfare would not change, while the value all other users derive from the resource would increase due to the decrease in congestion.

In any efficient allocation, it must be the case that for any user i and non-user j: $V_i - \beta_i g \geq V_j - \beta_j g$. This inequality implies that the value a user derives from use of the resource must be greater than the value that a non-user would derive from its use at the same congestion level. Otherwise, the respective user and non-user can

be interchanged, maintaining the same congestion level, and the total value from use of the resource would increase, violating the efficiency of the allocation. Users in an efficient allocation with congestion level g must then be those above and to the left of a ray with slope g, as in Figures 2.1 or 2.2. Additionally, the level at which the ray crosses the vertical axis must be the same as the access fee that would generate a level of congestion g in equilibrium. Otherwise, the congestion level and the number of users of the efficient allocation would not be compatible (the congestion level would not be that caused by the number of users).

The previous reasoning shows that the efficient allocation can be achieved by imposing an access fee at the adequate level. The optimization problem for the social planner can then be stated as finding the optimal access fee to maximize the economic value created by the resource.[6] Since this value is simply the sum of the values derived by users $(V_i - \beta_i g)$, we can express the optimization problem mathematically as

$$\max_{\tau} \quad \int_0^{\infty} \int_{\tau+\beta g(x(\tau))}^{\infty} [V - \beta g(x(\tau))] f(\beta, V) \, \mathrm{d}V \, \mathrm{d}\beta \tag{2.3}$$

where the function $x(\tau)$ (the number of users as a function of the access fee) is implicitly defined by Equation 2.2.

Appendix A shows that the first-order condition for this maximization problem gives the following formula for the optimal access fee τ^*

$$\tau^* = \bar{\beta} x^* g'(x^*), \tag{2.4}$$

where x^* is the optimal number of users, g' is the derivative of the congestion function

[6] The optimization problem could also be stated as choosing the number of users or the desired congestion level.

and $\bar{\beta}$ is the average sensitivity to congestion of users, which is given by

$$\bar{\beta} = \frac{\int_0^\infty \int_{\tau^*+\beta g(x^*)}^\infty \beta f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta}{\int_0^\infty \int_{\tau^*+\beta g(x^*)}^\infty f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta}. \tag{2.5}$$

Expression 2.4 can be interpreted as the traditional condition for a Pigouvian tax, which equates the tax to the marginal external cost of increased quantity. In this case, the marginal external cost is the product of the marginal effect on congestion of an additional user ($g'(x^*)$), the number of users (x^*) and their average sensitivity to congestion ($\bar{\beta}$).

2.2.4 Revenue-maximizing access fee

A private firm in charge of the congestible resource would like to impose an access fee to maximize its revenue, which equals the access fee times the number of users.[7] I assume the firm cannot price-discriminate, so it charges a unique fee τ to all users.

The problem for the private firm can be stated as

$$\max_{\tau} \quad \tau \cdot x(\tau), \tag{2.6}$$

where the function $x(\tau)$ is again implicitly defined by Equation 2.2.

Appendix A shows that the first-order condition for this maximization problem can be expressed as

$$\tau^p = \underbrace{\bar{\beta}_m x^p g'(x^p)}_{\substack{\text{Marginal congestion cost valued} \\ \text{according to marginal users}}} + \underbrace{\frac{\tau^p}{\bar{\varepsilon}}}_{\substack{\text{Monopolist} \\ \text{markup}}}, \tag{2.7}$$

where τ^p is the revenue-maximizing fee, x^p is the equilibrium number of users, $\bar{\beta}_m$ is

[7] I assume any maintenance costs of the facility are either zero or fixed, so they do not affect the pricing decision of the firm.

the average sensitivity to congestion of marginal users (those for whom $V_i - \beta_i g(x^p) = \tau^p$), which is given by

$$\bar{\beta}_m = \frac{\int_0^\infty \beta f(\beta, \tau^p + \beta g(x^p))\, \mathrm{d}\beta}{\int_0^\infty f(\beta, \tau^p + \beta g(x^p))\, \mathrm{d}\beta}, \tag{2.8}$$

and $\bar{\varepsilon}$ is the elasticity of demand with respect to the access fee when the access fee is τ^p and the level of congestion is held constant at $g(x^p)$.[8]

The first term in the righ-hand side of Equation 2.7 is similar to the expression of the optimal access fee (Equation 2.4). This term shows that the private firm internalizes the external congestion cost that an additional user imposes on other users. However, the private firm values this cost according to the average sensitivity to congestion of *marginal* users, while the optimal fee considers the average sensitivity of all users. This type of distortion between the incentives of a social planner and a monopolist was first highlighted by Spence (1975) with respect to the selection of product quality by a monopolist. It usually appears in contexts in which externalities affect heterogeneous users. For instance, Weyl (2010) identifies the same distortion in the case of network externalities in multi-sided platforms. Following Weyl, I will refer to this distortion as Spence distortion.

The second term in the right-hand side of Equation 2.7 can be interpreted as the usual markup imposed by a monopolist. It is inversely proportional to the elasticity of demand ($\bar{\varepsilon}$) when congestion is held fixed at $\bar{g} = g(x^p)$. If congestion were fixed at $\bar{g}$ (so the reservation value of each potential user equals $V_i - \beta_i \bar{g}$), this term would determine entirely the price imposed by the private firm.

[8]Note that this elasticity keeps the level of congestion fixed, so it does not correspond to the elasticity that would occur naturally as a results of price changes. Demand is more elastic when congestion is held constant because adjustments in the level of congestion tend to contain changes in the number of users.

2.2.5 Optimal vs revenue-maximizing fees

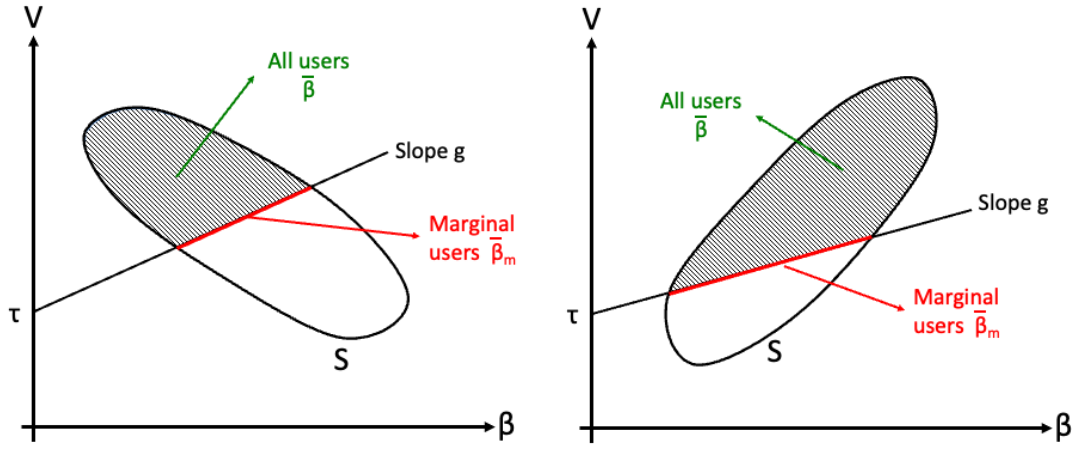
Equations 2.4 and 2.7 highlight two distortions between the welfare- and revenue-maximizing access fees: Spence and markup. According to the Spence distortion, the private firm can over- or under-internalize the external cost of congestion, depending on whether the average sensitivity to congestion of marginal users ($\bar{\beta}_m$) is higher or lower than that of all users ($\bar{\beta}$). The markup distortion reflects the usual tendency of a monopolist to raise prices above competitive levels to maximize profit.

The correlation between reservation values and sensitivities to congestion across potential users affects both distortions. First, a negative correlation causes $\bar{\beta} < \bar{\beta}_m$, while a positive correlation generates the opposite result. Figure 2.3 illustrates these two cases. A negative correlation then implies that the private firm over-internalizes the external congestion cost, while a positive one causes the firm to under-internalize it. Second, a positive correlation generates more similar reservations values at any given congestion level ($V_i - \beta_i \bar{g}$), because the burden of congestion is higher precisely for persons with high reservation values. As a result, the demand elasticity ($\bar{\varepsilon}$) is larger at any congestion level and the size of the markup decreases under positive correlation.

Negative correlation then simultaneously causes the private firm to over-internalize congestion and increases the size of the markup. As a result, the revenue-maximizing fee is likely to be well above the optimal one. Private ownership of the congestible resource would then result in a very inefficient low level of use. On the other hand, if the correlation is positive, the firm under-internalizes congestion and the Spence distortion mitigates, or maybe even outweighs, the markup. Private ownership is then more likely to achieve high efficiency levels under positive correlation.

Note that the previous analysis considers the distortions between welfare- and revenue-maximizing pricing *at a given level of use of the resource*. Equations 2.4 and 2.7 must be evaluated at the welfare- and revenue-maximizing number of users (x^*

Figure 2.3: Effect of correlation between reservation values (V) and sensitivities to congestion (β) on the relative sizes of the average sensitivity to congestion of all users ($\bar{\beta}$) and of marginal users ($\bar{\beta}_m$).



(a) Negative correlation ($\bar{\beta} < \bar{\beta}_m$). (b) Positive correlation ($\bar{\beta} > \bar{\beta}_m$).

Notes: τ denotes the access fee, g the congestion level and S the support of the bivariate distribution.

and x^p) respectively, and differences between these two levels will usually introduce additional distortions.[9] For example, even if there is no heterogeneity in sensitivities to congestion across potential users (no Spence distortion), the external cost of congestion considered by the private firm (first term in Equation 2.7) will differ from that considered by the social planner (Equation 2.4) due to differences in the number of users ($x^* \neq x^p$). Nevertheless, the Spence and markup distortions offer valuable intuition about the effect of correlation on the efficiency of private pricing. This intuition is confirmed in Section 2.3 through a numerical example borrowed from the transportation literature.

The importance of the correlation between reservation values and sensitivities to congestion for the comparison of the revenue- and welfare-maximizing access fees on congestible resources was first highlighted by Mills (1981). Mills employed a model of congestible resource use based on an inverse demand function that depends on the

[9]Tan and Wright (2018) wrote a comment to Weyl (2010) highlighting the importance of these additional distortions. Weyl inadvertently assumed that the welfare- and profit-maximizing conditions are evaluated at the same number of users.

number of users and the level of congestion: $\pi(x, g)$. This function indicates the most that would be paid by the xth user if the level of congestion were g. This approach is in fact more general than the model developed in this paper, because it does not assume that the level of congestion affects reservation values linearly. However, the model in this paper has several advantages over the one used by Mills. First, it clarifies the insights offered by Mills by explicitly considering the distribution of reservation values and sensitivities to congestion in the population of potential users. Second, it extends naturally to the analysis of several congestible alternatives. As explained in Section 2.2.7, users self-select into different congestible alternatives based on their sensitivties to congestion, so it is crucial to have an explicit distribution of sensitivities across the population. Finally, the bivariate distribution used in the model can be estimated empirically in a flexible manner,[10] while particular specifications of the inverse demand function employed by Mills would likely impose strong assumptions. For example, if the inverse demand function is linear on the number of users and price (i.e. $\pi(x, g) = a - bx - cg$), it implies that the distribution of reservation values across potential users is uniform at any given sensitivity to congestion.

It is reasonable to expect a positive correlation between reservation values and sensitivities to congestion in most applications, because the marginal utility of income usually plays an important role in both values. High income individuals are likely to have both high reservation values and high sensitivities to congestion. Consequently, studies that assume no correlation between these parameters (or assume a unique sensitivity to congestion across all users) are likely to overstate the inefficiency of private firms at managing congestible resources. On the other extreme, some studies assume perfect correlation between reservation values and sensitivities to congestion (e.g. Boffa et al. (2020)), which probably produces an overly optimistic view of private ownership. The following subsection analyzes the case of perfect correlation in more

[10] I estimate empirically this correlation for peak-hour travelers in Bogotá, Colombia, in Chapter 3.

detail.

2.2.6 The scale-income model

In his analysis of multi-sided platforms, Weyl (2010) recognizes the importance of heterogeneity in two dimensions for the comparison of profit- and welfare-maximizing pricing, obtaining formulas very similar to the ones presented in the previous sections. Weyl also recognizes that it may be hard to identify a two-dimensional model, so he proposes a model with heterogeneity in both dimensions but with perfect correlation between them, which in practice reduces heterogeneity to a single dimension. He named this approach the scale-income model, because it emerges from the idea that only differences in income levels (marginal utility of income) cause heterogeneity across potential users.

A simple way in which reservation values and sensitivities to congestion may be perfectly correlated is when the best outside option of all potential users involves the use of a similar alternative with a fixed level of congestion (for example, the ample road considered by Pigou and Knight in their old debate). If the level of congestion of the outside alternative is g_a, the reservation value of potential user i would be $\beta_i \cdot g_a$. In this case, potential users differ only in their sensitivity to congestion (β_i), but this heterogeneity also induces differences in reservation values. In terms of the bivariate distribution $f(\beta, V)$, the support of the distribution becomes a straight line.

According to the analysis of the previous section, perfect correlation between reservation values and sensitivities to congestion creates a very positive scenario for private ownership, so the scale-income model proposed by Weyl may not be a sensible choice to evaluate the efficiency of private ownership of a congestible resource (or a multi-sided platform). In fact, this approach can lead to the conclusion that private ownership achieves perfect efficiency. Take the case in which potential users choose between a congestible resource and an outside option with constant congestion level,

so user heterogeneity can be represented by a univariate distribution of sensitivities to congestion. Appendix B shows that in this case the revenue-maximizing fee equals the optimal fee if sensitivities to congestion follow a Pareto distribution. Interestingly, earnings and wealth distributions usually follow this type of thick-tailed distribution (Benhabib & Bisin, 2018).

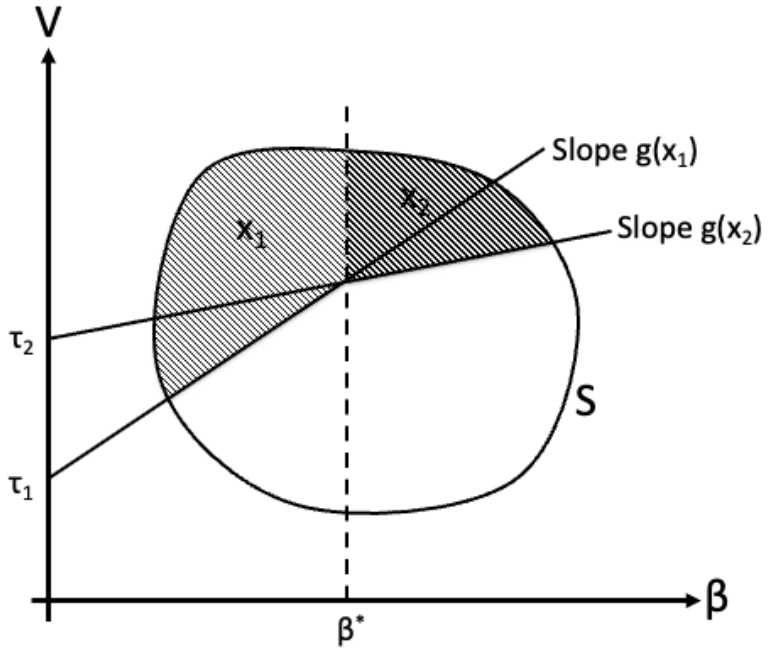
2.2.7 Two congestible resources

Let us now turn to the case in which potential users can choose between two congestible resources and an outside option, whose value is again normalized at zero. I assume the two resources are identical, so they have the same congestion function $g(x)$ and each potential user i assigns them the same reservation value V_i and sensitivity to congestion β_i. We can still use a bivariate density function $f(\beta, V)$ to define the distribution of preferences across potential users. Denote by x_1 and x_2 the number of users of resources 1 and 2. The value that user i derives from using resource 1 is $V_i - \beta_i g(x_1)$, while the value from resource 2 is $V_i - \beta_i g(x_2)$. Given access fees τ_1 and τ_2, user i prefers resource 1 over resource 2 if $\tau_2 + \beta_i g(x_2) > \tau_1 + \beta_i g(x_1)$.

Assuming $\tau_2 > \tau_1$, Figure 2.4 illustrates a Nash Equilibrium allocation of users between the two resources. The resource with the lowest access fee attracts more users and consequently reaches a higher congestion level ($x_1 > x_2$ and $g(x_1) > g(x_2)$). Users self-select between the two resources based on their sensitivity to congestion. Users with low sensitivities choose resource 1, while users with high sensitivities choose resource 2. β^* denotes the sensitivity threshold that separates the two groups.

We can then write the equilibrium conditions that determine the number of users

Figure 2.4: Nash Equilibrium of users for two identical resources.



Note: S represents the support of the bivariate distribution of sensitivities to congestion (β) and reservation values (V) across potential users. x_1 and x_2 identify the number of users of resource 1 and 2 respectively. τ_1 and τ_2 denote the access fee for each resource, while $g(x_1)$ and $g(x_2)$ represent the congestion level of each resource (a function of the number of users).

of each resource as a function of the access fees as

$$N \int_0^{\beta^*} \int_{\tau_1+\beta g(x_1)}^{\infty} f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta = x_1 \tag{2.9}$$

$$N \int_{\beta^*}^{\infty} \int_{\tau_2+\beta g(x_2)}^{\infty} f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta = x_2 \tag{2.10}$$

where the following equation defines β^*

$$\tau_1 + \beta^* g(x_1) = \tau_2 + \beta^* g(x_2) \tag{2.11}$$

The tolls on the two resources can be set by a social planner, a single private firm that manages both resources, or two competing private firms. The cases of a social planner and a single private firm lead to nice analytical conditions for the tolls, which are simple extensions of the conditions for the one-resource case. When two

firms compete (duopoly), it is not possible to obtain nice analytical conditions for the tolls. Even though the firms compete in prices (Bertrand competition) and there is no inherent differentiation between the two resources, the levels of congestion are an endogenous source of differentiation that leads to positive equilibrium prices. The numerical example in the next section shows that there are two types of equilibrium possible in this duopoly setting, but only one of them is stable.

2.3 Numerical example

This section explores quantitatively the effect of varying degrees of correlation between reservation values and sensitivities to congestion on the efficiency of alternative ownership arrangements of congestible resources. I borrow the numerical example developed by Verhoef and Small (2004). They studied road pricing under revenue- and welfare-maximizing objectives. Importantly, they used a continuous distribution of values of time along with linear inverse demand curves at each value of time. This demand specification maps into a specific bivariate distribution of reservation values and values of time $f(\beta, V)$ across potential travelers. Figure 2.5 graphs this distribution.[11]

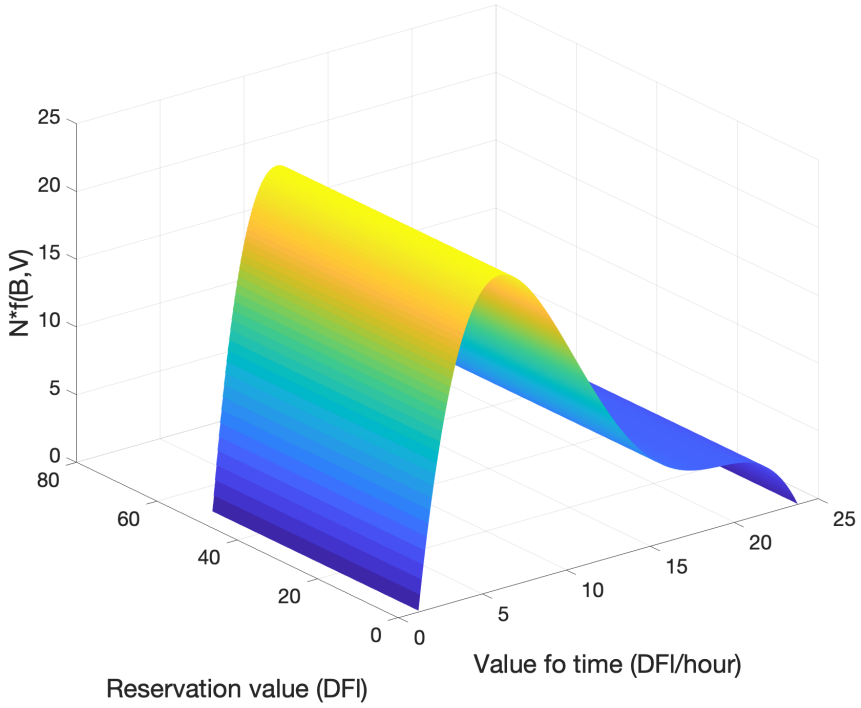
The starting point to arrive at this bivariate distribution was the univariate distribution of values of time shown in Panel 2.5c. This distribution of values of time was estimated in an earlier study for the Dutch Randstad area by fitting a fourth-order polynomial to stated-preferences data (Verhoef, Nijkamp, & Rietveld, 1997). Values of time vary from 1.2 to 23.8 DFl per hour. The assumption of linear inverse demand curves implies that the distribution of reservation values is uniform at each value of time. Verhoef and Small chose the support of the distribution and the total number of potential travelers ($N = 14,924$) to achieve a demand elasticity of -0.4 and a travel

[11] The density function of this bivariate distribution multiplied by the total number of potential travelers is

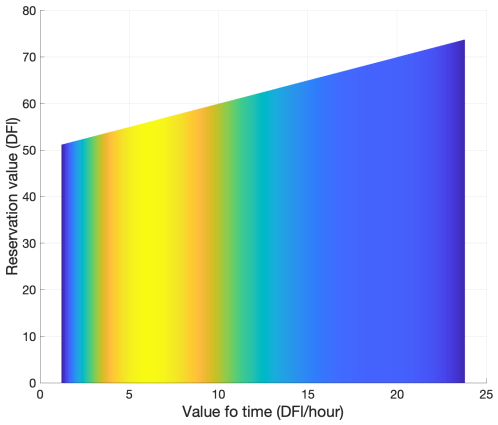
$$Nf(\beta, V) = (-0.713714 + 0.705429\beta - 0.0950357\beta^2 + 0.00468093\beta^3 - 0.000079\beta^4)/0.0434783$$

with the support shown in Panel 2.5b, whose upper limit is given by $V = 50 + \beta$.

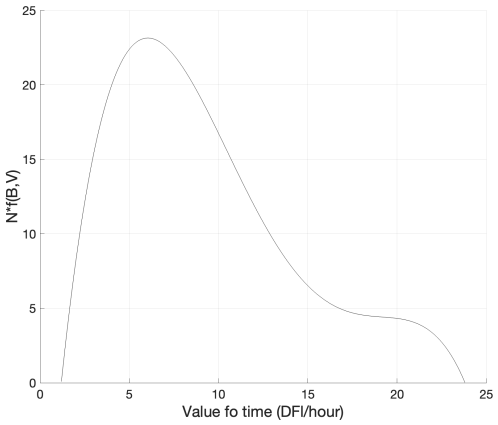
Figure 2.5: Bivariate distribution of reservation values and values of time from Verhoef and Small (2004).



(a) 3D view.



(b) Top view - Support of the distribution.



(c) Distribution of values of time.

Notes: The vertical axis in panels a and c corresponds to the bivariate density function times the number of potential travelers ($Nf(\beta, V)$).

time under free access equal to twice the free-flow travel time.[12]

The upper limit of the uniform distribution of reservation values increases slightly with value of time (see Panel 2.5b). This characteristic introduces a slight positive correlation between reservation values and values of time, and it implies the marginal distribution of reservation values is not exactly uniform and the marginal distribution of values of time does not exactly follow the fourth-order polynomial shown in Panel 2.5c. Since the procedure I will use to evaluate the impact of correlation takes as starting point these marginal distributions, I will slightly modify the support of the bivariate distribution to make it a rectangle, which simplifies the marginal distributions to uniform and fourth-order polynomial shapes. In the rectangular support, reservation values vary uniformly between 0 and 60 DFl at any value of time, while values of time vary between 1.2 and 23.8 DFL per hour at all reservation values. There is no correlation between these two parameters in the population. This adjustment to the distribution affects only slightly the main results obtained by Verhoef and Small, which are described below.[13]

2.3.1 One resource

Verhoef and Small (2004) were interested in the effect of second-best restrictions on the optimal and revenue-maximizing tolls, so they studied a road network with parallel and serial links, and with pricing possible only in some of the links. I am not interested in second-best restrictions, so I will employ a one-link representation of their network. Travel time (t, in hours) as a function of the number of travelers (x) is given by Equation 2.14. This functional form has been used extensively to study road congestion (Small & Verhoef, 2007, Section 3.3). The specific parameters of the function are meant to represent a four-lane highway with free-flow travel time of 0.5

[12] Verhoef and Small measure the elasticity of demand at the free access equilibrium and with respect to full price, which includes the cost of gasoline at 12 DFl per trip.

[13] The total number of potential travelers increases slightly to 15,123 due to the adjustment in the support of the distribution.

hours and capacity of 8,000 vehicles per hour.[14]

$$t(x) = 0.5\left(1 + 0.15\frac{x}{8,000}\right)^4 \tag{2.12}$$

As explained before, the bivariate distribution of reservation values and values of time used by Verhoef and Small, adjusted to a rectangular support, implies no correlation between these two parameters, while the marginal distributions of reservation values and values of time are uniform and fourth-order-polynomial respectively. In order to introduce different degrees of correlation to this bivariate distribution while maintaining the marginal distributions, I will use bivariate Gaussian copulas. Bivariate copulas describe the dependence structure between two random variables.[15] Bivariate Gaussian copulas describe the dependence structure of the bivariate normal distribution with varying coefficients of linear correlation (ρ). I use bivariate Gaussian copulas to generate bivariate distributions of reservation values and values of time that maintain the original marginal distributions but allow different degrees of correlation.[16]

Table 2.1 presents the numerical results of the model under free access, optimal pricing and revenue-maximizing pricing, for varying degrees of correlation.

The results of the base scenario, in which there is no correlation ($\rho = 0$), are close

[14] As Verhoef and Small (2004) clarify, this type of congestion function does not have a maximum flow, so the parameter of 8,000 vehicles per hour may be better interpreted as relative capacity.

[15] A copula is a multivariate cumulative distribution function for which the marginal distributions are uniform on the interval $[0, 1]$. The copula of a multivariate distribution can be obtained by applying the probability integral transformation to each component. By Sklar's theorem, any multivariate distribution can be expressed in terms of its marginals and copula.

[16] For a given coefficient of linear correlation (ρ) of the Gaussian copula, the density function of the bivariate distribution multiplied by the total number of potential travelers is

$$\begin{aligned} Nf(\beta, V; \rho) = {} & c(F_\beta(\beta), F_V(V); \rho) * \\ & (-0.713714 + 0.705429\beta - 0.0950357\beta^2 + 0.00468093\beta^3 - 0.000079\beta^4)/0.0434783 \end{aligned} \tag{2.13}$$

where $c(\cdot, \cdot; \rho)$ is the density function of the bivariate Gaussian copula with coefficient of correlation ρ, while $F_\beta(\cdot)$ and $F_V(\cdot)$ are the marginal cumulative distributions of reservation values and values of time. Note that the bivariate distribution $f(\beta, V; \rho)$ does not retain the coefficient of linear correlation ρ from the Gaussian copula because the integral transformation is not linear. Nevertheless, the linear correlation of the bivariate distribution does vary monotonically with ρ.

Table 2.1: Results under different pricing regimes and degrees of correlation.

	$\rho = -0.9$			$\rho = -0.4$			$\rho = 0$		
	Free	Optimal	Rev. Max.	Free	Optimal	Rev. Max.	Free	Optimal	Rev. Max.
Toll (DFl)	0	7.44	29.70	0	8.47	28.87	0	9.35	28.23
Travelers	12,177	11,028	6,602	12,499	11,003	6,666	12,816	10,985	6,759
Travel time (hours)	0.903	0.771	0.535	0.947	0.768	0.536	0.994	0.767	0.538
Revenue (DFl)	0	82,070	196,120	0	93,210	192,470	0	102,760	190,780
Welfare (DFl)	355,130	361,640	291,030	341,270	351,690	285,810	328,930	344,010	282,330
Rel. Efficiency	0	1	-9.84	0	1	-5.32	0	1	-3.09
$\bar{\beta}_m xt'(x)$ (DFl)			1.06			1.25			1.41
Markup (DFl)			28.64			27.62			26.82

	$\rho = 0.4$			$\rho = 0.9$		
	Free	Optimal	Rev. Max.	Free	Optimal	Rev. Max.
Toll (DFl)	0	10.31	27.62	0	11.72	26,92
Travelers	13,214	10,972	6,884	13,970	10,992	7,126
Travel time (hours)	1.058	0.765	0.541	1.197	0.767	0.547
Revenue (DFl)	0	113,090	190,120	0	128,810	191,860
Welfare (DFl)	314,260	336,400	279,210	288,660	327,120	276,300
Rel. Efficiency	0	1	-1.58	0	1	-0.32
$\bar{\beta}_m xt'(x)$ (DFl)			1.55			1.67
Markup (DFl)			26.06			25.25

Note: ρ determines the coefficient of linear correlation of the bivariate Gaussian copula.

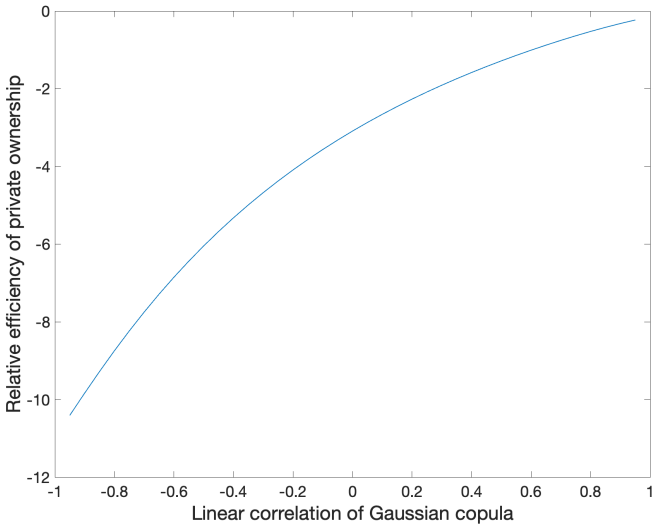
to those obtained by Verhoef and Small.[17] The difference arises from the adjustment of the bivariate distribution to a rectangular support. These results are very discouraging for private ownership. The revenue-maximizing toll is about three times higher than the optimal toll. As a result, the welfare created by private ownership is even lower than that achieved under free access. The last two rows of Table 2.1 dissect the revenue-maximizing toll into the two components introduced in Equation 2.7. The magnitude of the private fee is mainly determined by the size of the markup. Even though the lack of correlation in the bivariate distribution implies a small Spence distortion, the external congestion cost that the firm internalizes ($\bar{\beta}_m x t'(x)$) is significantly lower than the marginal external cost at the optimal level of use (which equals the optimal toll) because the number of travelers and consequently the slope of the congestion function are smaller under the revenue-maximizing fee.

The other scenarios in Table 2.1 introduce different degrees of correlation. The results show that the efficiency of private ownership varies considerably. The coefficient of relative efficiency in Table 2.1 normalizes the welfare achieved under the three pricing regimes to zero under free access and one under optimal pricing. The relative efficiency of private ownership varies from -9.84 in the scenario with strong negative correlation ($\rho = -0.9$) to -0.32 in the scenario with strong positive correlation ($\rho = 0.9$). Figure 2.6 graphs this relationship.

The main objective of Figure 2.6 is to highlight the strong effect that correlation has on the efficiency of private ownership. Even though Figure 2.6 also suggests that private ownership is always worse than free access (the coefficient of relative efficiency is always negative), this result cannot be generalized because it depends on the specific characteristics of the problem at hand. As an example, Figure 2.7 replicates Figure 2.6 using a different marginal distribution of reservation values. Verhoef and Small used a uniform distribution. I propose a log-normal distribution

[17] The three pricing regimes studied here correspond to the regimes that Verhoef and Small call No Toll (free access), Second-best Serial Link (optimal) and Private Serial Link (Rev. Max.).

Figure 2.6: Relative efficiency of private ownership as a function of the degree of correlation.



Note: The coefficient of relative efficiency normalizes the welfare achieved under free access to zero and under optimal pricing to one.

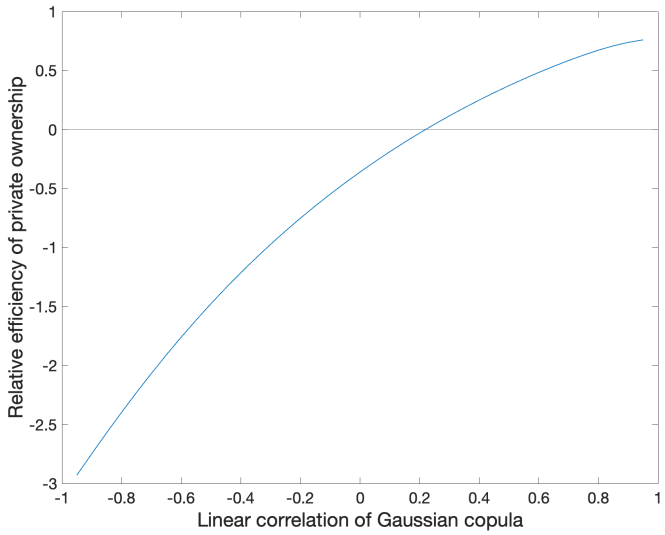
that maintains the same mean, variance and support of the uniform distribution.[18] Figure 2.7 shows that in this case private ownership is superior to free access as long as the degree of correlation is at least moderately positive ($\rho > 0.2$). Under strong positive correlations, private ownership achieves most of the welfare gains available from optimal pricing.

2.3.2 Two resources

To analyze the case of two identical congestible resources, I will divide evenly the relative capacity of the road. The example in this section then represents two two-lane roads with free-flow travel time of 0.5 hours and capacity of 4,000 vehicles per

[18]The log-normal shape is probably a more plausible representation of the distribution of reservation values, because the influence of income usually causes a long right tail. The uniform distribution had a support from 0 to 60 DFl, which implies a mean of 30 DFl and a variance of 300 DFl^2. To achieve the same mean and variance, the log-normal distribution has parameters $\mu = 3.257$ and $\sigma = 0.536$. I truncate the log-normal distribution at 60 DFl to maintain the same support.

Figure 2.7: Relative efficiency of private ownership as a function of the degree of correlation (log-normal distribution of reservation values).



hour. The travel time function of each road is

$$t(x) = 0.5\left(1 + 0.15\frac{x}{4,000}\right)^4 \tag{2.14}$$

The bivariate distribution of reservation values and values of time in the population of potential travelers is still the one used by Verhoef and Small adjusted to a rectangular support, with Gaussian copulas to add varying degrees of correlation.

I will consider four pricing regimes: free access (no toll on either road), optimal pricing (a social planner manages both roads), one private firm managing both roads, and two private firms each managing one road (duopoly). The results under free access are equivalent to those with one road (Table 2.1), with the number of travelers evenly divided between the two roads. Importantly, the total welfare achieved under free access remains the same. Table 2.2 presents the results of the other three pricing regimes for varying degrees of correlation. The coefficient of relative efficiency is still normalized to zero for free access and one for optimal pricing.

Looking at the base scenario ($\rho = 0$), both the social planner and the single

Table 2.2: Results for two roads under different pricing regimes and degrees of correlation.

		$\rho = -0.9$			$\rho = -0.4$			$\rho = 0$		
		Optimal	One firm	Duopoly	Optimal	One firm	Duopoly	Optimal	One firm	Duopoly
Road 1	Toll (DFl)	6.77	29.63	10.56	7.52	28.76	11.24	8.25	28.08	12.05
	Travelers	5,936	3,535	5,370	6,025	3,630	5,361	6,045	3,691	5,336
	Time (hours)	0.864	0.546	0.744	0.886	0.551	0.742	0.891	0.554	0.737
	Revenue (DFl)	40,200	104,760	56,730	45,290	104,410	60,240	49,870	103,670	64,290
Road 2	Toll (DFl)	7.86	29.74	10.85	8.97	28.93	11.55	9.91	28.31	12.40
	Travelers	5,156	3,075	5,109	5,071	3,050	5,094	5,052	3,083	5,067
	Time (hours)	0.707	0.526	0.700	0.694	0.525	0.697	0.691	0.526	0.693
	Revenue (DFl)	40,530	91,450	55,440	45,470	88,210	58,860	50,080	87,300	62,830
Total Revenue (DFl)		80,730	196,210	112,170	90,760	192,620	119,100	99,950	190,970	127,120
Total Welfare (DFl)		362,660	291,310	360,740	353,370	286,290	351,080	346,030	282,910	343,390
Rel. Efficiency		1	-8.48	0.74	1	-4.54	0.81	1	-2.69	0.85

		$\rho = 0.4$			$\rho = 0.9$		
		Optimal	One firm	Duopoly	Optimal	One firm	Duopoly
Road 1	Toll (DFl)	9.16	27.44	12.99	10.77	26.68	14.32
	Travelers	6,026	3,753	5,305	5,942	3,887	5,304
	Travel time (hours)	0.886	0.558	0.732	0.865	0.567	0.732
	Revenue (DFl)	55,170	103,000	68,930	63,980	103,710	75,950
Road 2	Toll (DFl)	10.93	27.74	13.38	12.40	27.10	14.82
	Travelers	5,067	3,148	5,036	5,155	3,260	5,003
	Travel time (hours)	0.693	0.529	0.688	0.707	0.533	0.684
	Revenue (DFl)	55,400	87,320	67,400	63,920	88,370	74,160
Revenue (DFl)		110,570	190,320	136,330	127,900	192,080	150,110
Welfare (DFl)		338,480	279,840	335,610	328,700	276,920	326,050
Rel. Efficiency		1	-1.42	0.88	1	-0.29	0.93

Notes: ρ determines the coefficient of linear correlation of the bivariate Gaussian copula.

firm take advantage of the split in capacity to differentiate the two roads. One of the roads gets a lower toll (by notation Road 1) and attracts more travelers, which causes a higher congestion level. Travelers with low values of time self select into this road. The social planner, however, imposes a larger differentiation between roads, as measured by the difference between tolls. The welfare achieved by the social planner and the revenue achieved by the private firm increase slightly in comparison to those achieved with only one road, which reveals that they both benefit from the split in capacity.[19] The tolls imposed by the private firm are still well above the optimal ones, so the welfare achieved by a single firm in charge of both roads is still relatively low.

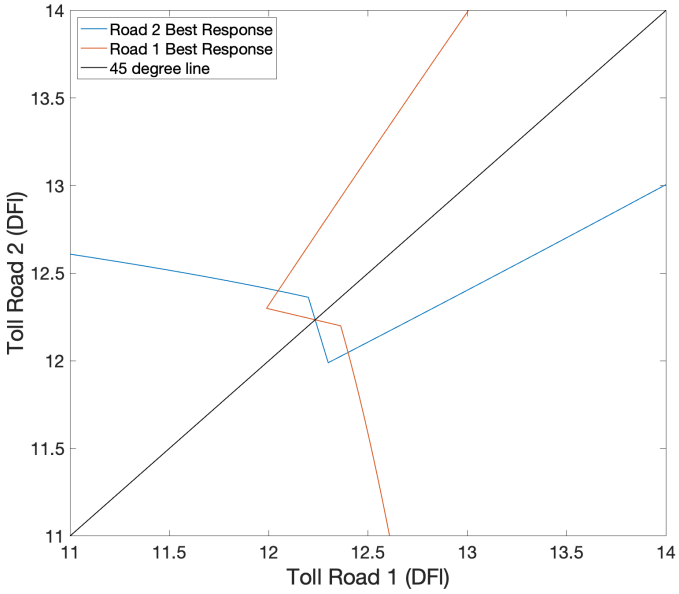
When two private firms manage one road each (duopoly), their competition drives prices down. As a result, the welfare achieved in this regime is much higher than with a single firm, and close to the highest possible (coefficient of relative efficiency 0.85). The differentiation between roads in this regime is still low (close to the differentiation imposed by a single firm), and the revenue achieved by the two firms is different but similar (about 2.3% difference). Interestingly, the firm that charges the lower toll receives the higher revenue.

The pricing game between the two firms does not have a unique equilibrium. Figure 2.8 graphs their best response functions.[20] There are three equilibria. One of these equilibria has the two firms setting the same toll (symmetric equilibrium). This equilibrium, however, is not stable (it cannot be reached by successive best responses by the firms). The other two equilibria are equivalent and achieve the same otucomes. They simply differ on which firm takes the low-toll role. These two equilibria are stable. I assume the firms reach one of these two equivalent equilibria.

[19]Note that the congestion function used in this example has constant returns to scale (i.e. it is homogeneous of degree zero with respect to the ratio between users and capacity), so there are no disadvantages from splitting capacity. If the congestion function had decreasing returns, the social planner and the private firm may not benefit from the split.

[20]Note that when the toll of the other firm is low, the best response by a firm to an increase in the toll by the other is to decrease its toll (strategic substitutes). But this pattern changes when the toll of the other firm is relatively high (strategic complements).

Figure 2.8: Best response functions of the tolling game for two private firms managing one road each.



The efficiency achieved by a single firm varies with the degree of correlation in a manner similar to the one-road case, where positive correlation leads to a more efficient result. Positive correlation also improves the relative efficiency of the duopoly results. However, the effect is weaker for this regime. The welfare achieved by a duopoly is always higher than under free access and close to the highest possible.

2.4 Conclusions

Since Pigou started the study of optimal pricing of congestible resources almost 100 years ago, the question of to what extent private pricing leads to an efficient level of use has stimulated economic research. This research is relevant to modern industries that feature congestion and market power, such as the internet and the electricity grid, as well as traditional common-pool resources, such as oil and fisheries. The economic literature currently recognizes two main distortions between optimal and private pricing. First, there is the usual markup that a monopolist imposes to max-

imize profit. Second, even though the monopolist internalizes the external effect of additional users on congestion, she values it according to the average sensitivity to congestion of *marginal* users, while the correct valuation should consider all users (Spence distortion). User heterogeneity, both in terms of reservation values and sensitivities to congestion, determines the size of these distortions. Additionally, the degree of correlation between these two dimensions of heterogeneity affects both distortions. Positive correlation reduces the size of the markup and causes the monopolist to under-value the external cost of congestion. As a result, positive correlation usually improves the efficiency of private pricing.

This chapter contributes to this line of research by analyzing a model of congestible resource use that explicitly considers the bivariate distribution of reservation values and sensitivities to congestion across potential users. From this model, I derive analytical conditions for the welfare- and revenue-maximizing access fees, which highlight the markup and Spence distortions and the importance of correlation for the efficiency of private pricing. The model extends easily to two or more congestible alternatives, where users self select into the alternatives according to their sensitivity to congestion. This extension allows for the study of a novel duopoly setting with endogenous congestion, no inherent differentiation between the two alternatives and Bertrand competition in prices. Even though this new duopoly setting offers two types of equilibrium, only one of them is stable (i.e. can be reached by successive best responses by the firms).

This chapter also explored quantitatively the impact of correlation using a numerical example borrowed from the road pricing literature, adding varying degrees of correlation through Gaussian copulas. The results show that the efficiency of private pricing varies considerably with the degree of correlation. Private pricing is several times worse than free access under negative correlation, while it can get close to complete efficiency under strong positive correlations. In fact, the extreme case of

perfect correlation between reservation values and sensitivities to congestion (named scale-income model by Weyl (2010)) leads to complete efficiency when income follows a Pareto distribution. Even though positive correlations also improve the efficiency of a duopoly, this effect is significantly more tenuous. A duopoly achieves high efficiency even in the presence of strong negative correlations.

It is common for studies that evaluate the efficiency of private pricing of congestible resources to assume a correlation between reservation values and sensitivities to congestion. The results of this paper suggest these studies may incur in significant bias. For instance, studies that assume no correlation would overestimate the inefficiency of private pricing if there exists in fact a positive correlation. Theoretical models of congestible resources should ideally be based on a general bivariate distribution of these two individual characteristics, while numerical studies should explore the sensitivity of their results to varying degrees of correlation. Even though the influence of income on both characteristics suggests that most real life situations involve strong positive correlations, the degree of correlation is ultimately an empirical question. Empirical studies that identify this correlation will greatly contribute to the body of knowledge about the efficiency of private pricing of congestible resources.[21]

[21] I estimate empirically this correlation for peak-hour travelers in Bogotá, Colombia, in Chapter 3.

Chapter 3

Congestion Charges Under Market Power: An application to ride-hailing in Bogotá, Colombia

3.1 Introduction

The rapid growth of ride-hailing services mediated by digital platforms like Uber, Lyft or DiDi over the last decade has caused concerns about their potential to worsen traffic congestion and other transportation-related externalities in cities worldwide. As a result, congestion charging (the adaptation of a Pigouvian tax to traffic congestion externalities) has been suggested as a mitigation strategy. For example, New York City implemented in February 2019 a congestion surcharge of $2.75 for ride-hailing trips entering or passing through Manhattan south of 96th street (New York City Taxi and Limousine Commission, 2019).[1]

Negative externalities, however, are not the only failure of ride-hailing markets. Market power is also prevalent. Even though there are usually thousands of riders and drivers, only one or two digital platforms control the market in each city. For example, Uber and Lyft control almost the entire market in most U.S. cities (Statista, 2020),

[1]New York is also expected to be the first U.S. city to introduce congestion pricing for private vehicles in 2021. Even though economists have long argued for congestion pricing as a tool to manage traffic congestion, only a handful of cities around the world have actually implemented it. The New York case suggests it may be easier to implement congestion pricing for ride hailing than for private cars. Technological barriers are clearly lower for ride hailing because platforms already have the technology in place to identify and charge individual trips, so no need for E-Z passes or cameras taking photos of license plates. Other cities that have implemented surcharges on ride hailing include Mexico City (1.5%), Chicago ($0.69 USD), Rio de Janeiro (1%), Calgary ($0.30 CAD) and San Francisco (3.25%) (Yanocha & Mason, 2019).

while DiDi and Grab are dominant in China and southeast Asia respectively. Ride-hailing markets tend to concentrate in very few platforms mainly because of network effects in wait times. A platform serving proportionally more riders and drivers can offer lower wait times to travelers (Frechette et al., 2019). In their typical business model, ride-hailing platforms not only match riders and drivers but also set prices to both sides of the market. Platforms can then exert market power to impose a profit-maximizing gap between these two prices (commonly known as platform commission), in the same manner a monopolist producer imposes a profit-maximizing markup.

Economists have long recognized that market structure can significantly influence the efficiency of Pigouvian taxes (Buchanan, 1969). This insight can be easily grasped. Picture a market monopolized by a producer, who imposes a profit-maximizing markup and causes an external cost (larger than the markup). A regulator concerned about the externality may wish to impose a Pigouvian tax equal to the marginal external cost (MEC), but this policy would push the price faced by consumers above the optimal level. The optimal tax to address the externality must be smaller than the MEC because there is a smaller gap to bridge. In a more extreme, but equally plausible, scenario, the markup may be larger than the MEC, in which case *any* positive tax would be detrimental. In a lucky situation, MEC=markup, the monopolist charges a socially efficient price and there is no need for government intervention to achieve efficiency.

In this chapter, I apply the previous observation to ride-hailing markets in order to judge the merit of a congestion charge. Since ride-hailing markets feature both externalities and market power, the optimal congestion charge is smaller than the MEC of congestion, and may actually turn negative if the markup is substantial. This chapter provides the first empirical comparison between congestion externalities and market power in the ride-hailing industry. To do so, I set up and calibrate for Bogotá, Colombia, a structural model of ride hailing, which ultimately allows me to

estimate the optimal charge under a monopolistic market structure.

The structural model has four components. The first one is a demand model for ride-hailing services. I propose and estimate empirically based on stated-preference surveys a demand model that allows for individual heterogeneity in reservation values and values of time. Importantly, the model allows also for correlation between these two dimensions of heterogeneity. This correlation, usually neglected in transportation demand models, can have a significant impact on the divergence between monopolistic and optimal price levels, as highlighted in Chapter 2. Travel time for ride hailing is the sum of wait time and in-vehicle time. Both of these time factors depend on the number of ride-hailing vehicles on the street. In-vehicle time increases with the number of vehicles due to traffic congestion. Wait time decreases with it because there is a better chance an idle vehicle is close to the rider's location. As a result, the number of travelers willing to hail a ride depends not only on the price charged by the platform, but also on the number of ride-hailing vehicles available for service.

The second component is a supply of drivers that adjusts to achieve a fixed revenue per hour. The assumption behind a constant revenue per hour is that all potential drivers have the same reservation wage, which includes vehicle operating expenses (e.g. gasoline) and net earnings. J. V. Hall, Horton, and Knoepfle (2020) and Alvarez and Argente (2020) provide empirical evidence suggesting this assumption is a good approximation to the labor supply of ride-hailing drivers. However, this assumption brings two important and interrelated implications, which should be kept in mind to ponder some of the results. First, no surplus is created on the side of drivers (all drivers earn their reservation wage). Second, the monopsonistic position of a platform in the market for ride-hailing drivers does not result in a markdown in the price paid to drivers. The uniform-reservation-wage assumption is equivalent to a constant marginal cost of production.

The third component is a matching process between riders and drivers, which de-

termines wait times. I assume platforms match riders to the closest idle vehicle, while idle vehicles are evenly distributed over the service area. These two assumptions determine a mathematical relationship, first derived by Arnott (1996), between average or expected wait time and the density of idle vehicles in the service area. The number of idle vehicles is, in steady state, a function of the total number of vehicles available for service, the number of trips requested per hour and average travel time. This function introduces network effects into the picture, because a proportional increase in riders and drivers (i.e. an increase in the platform's scale) raises the number of idle vehicles and consequently lowers wait times.

The fourth and last component of the structural model is an empirical estimate of the marginal effect of additional vehicles on average travel speed. This magnitude is clearly at the heart of traffic congestion externalities. I measure it for Bogotá based on data from its 2019 Mobility Survey and applying the methodology originally proposed by P. A. Akbar and Duranton (2017). In this methodology, the effect of additional vehicles on travel speed is identified from changes in traffic volume throughout the day, controlling for concurrent changes in trip and traveler characteristics.

These four components interact to determine the number of riders and drivers in equilibrium for a given set of prices per trip charged to riders and paid to drivers. Once prices are set, ride-hailing markets clear on travel time (in-vehicle plus wait).

I study three pricing regimes or scenarios. In the base scenario, a monopolistic platform sets prices to maximize profit. Profit equals the product between the number of riders and the price gap or platform commission. In an ideal scenario, a benevolent social planner controls the platform and sets prices to maximize total welfare. Total welfare includes the surplus created for ride hailers (there is no surplus for drivers), minus the cost of vehicles and drivers and the traffic congestion externality imposed on other road users. Finally, in an economist's dream, a private platform sets prices but a regulator is able to force socially efficient outcomes through taxation.

I derive analytical conditions for the profit- and welfare-maximizing price gaps. These conditions reproduce the contest between markup and marginal external cost that drives the sign and magnitude of the optimal congestion charge. Additionally, they reveal that a private platform internalizes (with a distortion) the congestion externality each ride hailer imposes on other ride hailers (not on other road users). As a result, the platform internalizes a larger portion of the total external cost of congestion as ride-hailing vehicles become a larger percentage of total traffic volume, potentially weakening the motivation for a congestion charge.

I calibrate the components of the structural model to the morning peak period of an average weekday in 2019 in Bogotá, Colombia. Bogotá is a highly dense and congested city of about 7.5 million inhabitants.[2] According to its 2019 Mobility Survey (Secretaría Distrital de Movilidad, Bogotá D.C., 2019), the number of ride-hailing trips per hour during the morning peak period was about 11,000, which is a high number but represents only 1.1% of all trips in the city.[3] The survey also reveals that Uber had a strong hold on the market with about 70% of all ride-hailing trips.

The numerical results for Bogotá indicate that, in a monopolistic scenario, the platform imposes a 18.5% gap or commission between the prices charged to riders and paid to drivers. This gap represents the markup imposed by the platform. The rider fare in this scenario is less than 5% larger than the average fare observed in the mobility survey, which suggests the monopolistic scenario is a good approximation to the situation of Bogotá's ride-hailing market in 2019. When a social planner manages the platform, the price gap increases to 25.9%. This optimal gap reflects mainly the external cost ride hailing imposes on other road users through traffic congestion. In this case, the marginal external cost turns out to be larger than the markup, so a

[2]Bogotá usually tops worldwide congestion rankings such as the Global Traffic Scorecard (Inrix, 2019).

[3]As a comparison, the number of ride-hailing trips per hour during the evening peak in San Francisco in 2016 was about 14,000, while ride-hailing trips accounted for 9% of all weekday trips in the city (San Francisco County Transportation Authority, 2017).

congestion charge on ride hailing is justified.

I estimate the optimal congestion charge at COL$1,700 for an average-distance trip during the morning peak period, which represents only 50% of the marginal external cost of congestion. This charge can be implemented as a 14.6% tax on riders' fare, or as a COL$222 charge per kilometer.[4] The optimal congestion charge takes into account an incomplete pass-through of the tax by the platform to riders. The model reveals a monopolistic pass-through of 0.81 (i.e. in response to a $1,000 tax, the platform lowers its price by about $190).

Even though regulators should in theory take care of both sides of the market (riders and drivers) in order to fully achieve socially efficient outcomes, the optimal tax on the side of riders does most of the job. It achieves about 98% of the welfare gains available from the unregulated scenario. As a result, there is little incentive to also regulate the side of drivers.

A few considerations about these results are in order. First, the charge does not account for other transportation-related externalities such as traffic accidents and air pollution, which would enlarge it.[5] Second, the optimal charge for other time periods is likely to vary significantly, especially for periods with low traffic volume. Finally, this chapter assumes regulators do not impose a congestion charge on private cars. Such a (desirable) policy would affect the optimal charge on ride hailing to the extent it reduces traffic volume or increases the demand for ride hailing.[6]

The remainder of this chapter contains four sections. Section 3.2 details the components of the structural model and presents analytical conditions for the profit-

[4]The optimal congestion charge in the model is proportional to trip distance, so it should be applied on a per-kilometer basis or as a percentage of the fare, not as a fixed amount to all trips.

[5]Congestion externalities are usually larger than accidents and air pollution (see for instance Parry and Small (2005)), especially during peak hours, so the optimal charge may not increase considerably due to these other externalities.

[6]Bogotá currently takes 50% of the private car fleet out of circulation during peak hours through a license plate-based restriction. This restriction probably contributes significantly to the demand for ride hailing, so changes to this policy would likely affect the optimal congestion charge on ride hailing.

and welfare-maximizing price gaps. Section 3.3 describes the data and empirical estimations carried out to calibrate the model to Bogotá. Section 3.4 reveals the main numerical results. I wrap up in Section 3.5 with the main conclusions. Before all that, I will describe how this chapter connects and contributes to several bodies of literature.

Related literature

The emergence of ride-hailing platforms over the last decade has inspired a rapidly growing literature with diverse emphases. One initial concern has been the potential of ride hailing to increase vehicle-miles traveled (VMT) in cities, consequently exacerbating traffic congestion and other transportation-related externalities. There was an initial debate in the transportation literature about whether ride hailing contributes or not to urban VMT. Platforms regularly argue that ride hailing can reduce VMT by facilitating access to public transportation, reducing the need to own cars or providing shared services. A few studies supported these arguments. For instance, J. D. Hall, Palsson, and Price (2018) found that Uber increased transit ridership by 5% in average two years after entry to U.S cities. However, the most recent literature with detailed data concludes that ride hailing does contribute to VMT and congestion, mainly because it replaces many trips that would have been made by more sustainable modes such as public transportation. For example, Erhardt et al. (2019) conclude that ride hailing had a significant effect on congestion in San Francisco between 2010 and 2016. Tirachini (2020) provides an international review including experiences from developing countries.

On the other hand, several studies have quantified the economic value created by ride-hailing platforms. Cohen, Hahn, Hall, Levitt, and Metcalfe (2016) and Lam and Liu (2017) estimate that riders gain a surplus of $1.60 and $0.72, respectively, per dollar spent on platforms in major U.S. cities. Frechette et al. (2019), Buchholz

(2020), Bian (2018) and Shapiro (2018) develop dynamic and spatial equilibrium models to estimate the efficiency gains that electronic matching offers over street hailing. Castillo, Knoepfle, and Weyl (2018) and Castillo (2019) focus on the welfare gains available from dynamic pricing. Finally, Chen, Chevalier, Rossi, and Oehlsen (2019) measure the value drivers derive from being able to choose when to work.

This chapter bridges these two opposing views of ride hailing. Even though digital platforms increased the efficiency of rider-driver matchings and implemented dynamic pricing, these improvements do not imply they raised overall welfare due to two factors. First, efficiency improvements may result in lower overall welfare in the presence of unregulated externalities such as traffic congestion. Second, the pricing strategies pursued by private platforms with market power may differ considerably from the socially optimal ones. Only through proper regulation we can extract the full benefits of technological improvements and guarantee they do not decrease welfare. The main objective of this chapter is to help design such regulation for the ride-hailing industry.

This chapter also contributes to the literature on environmental regulation under market power. Buchanan (1969) pointed out that environmental regulation designed to completely internalize external damages in non-competitive industries may reduce welfare. Fowlie et al. (2016) confirmed this possibility for the U.S. cement industry. They found that policies designed to internalize the social cost of carbon in this industry reduce welfare. Consequently, optimal carbon pricing involves firms only partially internalizing the social cost of carbon. My results offer a similar conclusion. The optimal congestion charge on ride hailing corresponds to only 40% of the marginal external cost of congestion.

Finally, this chapter contributes to the economic literature on monopolistic pricing of congestible resources. Mills (1981) showed that a monopolist internalizes, with a distortion, the congestion effects on users of its resource, and revealed that the

magnitude of this distortion depends on the correlation between reservation values and sensitivities to congestion (values of time in our context) in the population of potential users. Chapter 2 confirmed quantitatively the importance of this correlation for the efficiency of private pricing of congestible resources. Even though the transportation economics literature recognizes this effect (see for instance Brueckner (2002)), this paper presents the first empirical estimate of correlation between reservation values and values of time across a population of travelers.

3.2 Model and analytical solutions

This section details the four components of a model that intends to capture the main features of the ride-hailing industry as it has come to be in the last decade. Ride-hailing platforms electronically match riders and drivers, and set prices to both sides of the market.[7] Potential riders can check the price and estimates of in-vehicle and wait times on their smartphones before deciding to request a ride. On their side, drivers are free to choose when to be available for service, basing their decisions on previous experience about average earnings.[8] An economic model of ride hailing must then include at least three components: a demand model that represents travelers' alternatives, a supply model that represents drivers' decisions to work, and a matching process that determines wait times. Since one of the main objectives of this paper is to evaluate the impact of ride hailing on traffic congestion, I add a fourth component that endogenizes in-vehicle travel time as a function of the number of ride-hailing vehicles on the street.

I simplify platforms' pricing decisions down to one price to charge riders and one price to pay drivers for an average-distance trip. It is implicitly assumed that

[7] Other business models for platforms exist but have not become mainstream. For example, InDriver allows drivers to name their price for each trip while riders select the best offer. Empower does not impose a platform commission, but charges drivers a flat monthly fee.

[8] Drivers usually do not know the amount they will earn for, and have in general limited information about, a trip before accepting it. Therefore, their role in picking trips is small.

prices for diverse trips are proportional to distance, both in monopolistic and socially optimal scenarios.[9] Accordingly, the optimal congestion charge will be computed for an average-distance trip, but should be applied to diverse trips on a per-kilometer basis or as a percentage of the price charged by the private platform.

The previous simplification implies that the model does not account for potential differences between private and social incentives in price differentiation across space and time. In space, Bimpikis, Candogan, and Saban (2019) show that private platforms can benefit substantially from pricing rides differently depending on the location of origin. Although private and social pricing incentives coincide when the demand is balanced across locations, they will in general deviate for unbalanced demands, which are the norm for urban transportation. This deviation implies that optimal congestion charges should also vary according to the specific location of origin of the trip. The computation of optimal location-specific congestion charges, however, falls outside the scope of this chapter.

In time, private platforms may set prices for different time periods in an interdependent way if demand or supply are related across time periods. Demand may be related across time periods if travelers can adjust the start time of their trip in response to prices. Supply may also be related if drivers' decision to work in a time period depends on whether they work on an adjacent period (e.g. drivers may prefer to work consecutive rather than alternative periods). The model evaluates pricing within a single time period of the day, ignoring any potential interdependencies in pricing across different time periods. The time period of analysis must then be ample

[9]The model does not miss an important distortion between monopolistic and optimal pricing to the extent that prices are in fact roughly proportional to distance in both scenarios. Profit-maximizing prices may not be proportional to distance if some traveler's characteristics, such as income, are significantly correlated with distance, because platforms may then price-discriminate based on distance (e.g. price long trips relatively higher if they are mainly made by high-income travelers). However, the usual pricing strategy by platforms, which includes a fixed charge plus variable charges based on distance and travel time, implies that prices are approximately proportional to distance. Socially optimal prices may not be proportional to distance if, for example, long trips tend to add proportionally more or less to congestion than short ones.

enough to believe it would be too costly for travelers to adjust their start times to other periods, and to believe drivers can choose to work during this period without having to work on adjacent ones. For Bogotá, I analyze the morning peak period (6-8:30am).

3.2.1 Demand: Riders' side

The first time you install a ride-hailing app on your phone, you may decide to use it for a trip that you would have otherwise done by other means (e.g. by bus), or you may actually decide to travel because the app now makes it convenient. The population of potential ride hailers is then composed of people currently traveling by other modes as well as people not traveling. I will, nonetheless, most of the time refer to potential ride hailers as travelers to ease wording. I will refer to people actually requesting rides as rider or rider hailers.

Travelers choose between ride hailing and their best alternative or outside option (other mode or not traveling) within the time period of analysis.[10] Normalizing the value of the outside option to zero, the value traveler i gets from ride hailing can be expressed as:

$$\text{Ride hailing} : V_i - \beta_i \cdot [t(d) + w(x, d)] - p \tag{3.1}$$

$$\text{Outside option} : 0$$

V_i is the reservation value in monetary terms traveler i assigns to ride hailing in comparison to her outside option. V_i reflects individual preferences as well as the characteristics of outside options (e.g. price and travel time of alternative modes). β_i is the value of time of traveler i. It is a measure of how travel time lowers reservation values.[11] $t(d)$ is in-vehicle travel time as a function of the number of ride-hailing

[10] I assume people make at most one trip during the time period of analysis, which is a reasonable assumption, especially for peak periods.

[11] The transportation literature highlights that most travelers value wait time more than in-vehicle time (Small, 2012). I could allow these two valuations to vary independently across travelers, but that

vehicles on the streets. In-vehicle time increases with the number of vehicles due to traffic congestion (see Section 3.2.4). $w(x, d)$ denotes average or expected wait time as a function of the number of riders or trips requested per hour (x) and the number of vehicles (d).[12] Wait time increases in x and decreases in d (see Section 3.2.3). Finally, p is the price per trip charged to riders.

A reservation value (V_i) and a value of time (β_i) identify each traveler. A bivariate probability density function $f(\beta, V)$ can then represent the distribution of individual preferences across the population of potential ride hailers. For a given price (p), in-vehicle time (t) and wait time (w), traveler i chooses ride hailing if $V_i - \beta_i \cdot (t+w) - p > 0$. For a given set (p, t, w), riders can then be identified in a two-dimensional graph, with β in the horizontal axis and V in the vertical one, as those above and to the left of a ray that goes through $(0, p)$ and has slope $t + w$. Figure 3.1 illustrates this characterization, where S denotes the support of the bivariate distribution $f(\beta, V)$.[13]

Since riders raise wait times, the equilibrium number of riders for a given price and number of vehicles must solve the following equation (where N represents the size of the population of potential ride hailers):

$$N \int_{\beta=0}^{\infty} \int_{V=p+\beta(t(d)+w(x,d))}^{\infty} f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta = x \tag{3.2}$$

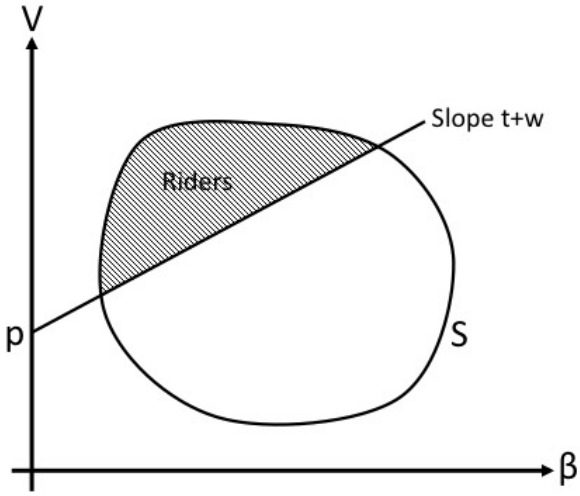
Equation 3.2 determines the number of riders (x) as an implicit demand function of price (p) and number of vehicles (d).

would introduce an additional dimension of individual heterogeneity, further complicating empirical estimation and numerical solutions.

[12] I assume one rider corresponds to one trip. The model does not consider shared rides.

[13] I assume the support lies entirely in the first quadrant (positive V and β). It would take a very strange individual to encounter a negative value of time. It is plausible for individuals to have negative reservation values, but their existence is immaterial for our purposes because they would never choose ride hailing, unless platforms decide to pay riders (negative price) or we figure out a way to travel back in time (negative travel time).

Figure 3.1: Riders for a given price (p), in-vehicle time (t) and wait time (w).



Note: S identifies the support of the distribution of reservation values (V) and values of time (β) in the population of potential riders.

3.2.2 Supply: Drivers' side

The side of drivers is simpler. Drivers choose to work during the time period of analysis if expected earnings are higher than their reservation wage. Since drivers are usually responsible for all vehicle expenses (e.g. gasoline consumption), their reservation wage includes these expenses plus net earnings. I assume all potential drivers have the same reservation wage, denoted c (for cost, w was already taken by wait time). This assumption may seem extreme, but J. V. Hall et al. (2020) provide empirical evidence suggesting it is a good approximation to the labor supply of ride-hailing drivers. They show that after Uber-initiated price increases in U.S. cities, driver supply adjusted to bring hourly earnings back to their initial level.[14] Similarly, Alvarez and Argente (2020) find evidence of a very elastic supply side for ride hailing in Mexico and Panama.

Expected hourly earnings for a driver equal the product of the number of trips per hour she expects to serve and the price per trip paid by the platform (q). Assuming trips are evenly distributed among drivers, the expected number of trips for a driver

[14]The uniform-reservation-wage assumption can also be considered an approximation to a large pool of potential drivers.

equals the ratio between total trips (x) and total drivers (d).

The equilibration process on the side of drivers is then straightforward. Drivers enter the market until expected earnings equal the common reservation wage (c). Mathematically:

$$c = \frac{x}{d} \cdot q \tag{3.3}$$

Equation 3.4 determines the number of drivers or vehicles (d) as a supply function of price (q) and number of riders per hour (x). Contrary to the demand function, the supply function is very explicit:

$$d = \frac{q}{c} \cdot x \tag{3.4}$$

The uniform-reservation-wage assumption brings two important and interrelated implications, which should be kept in mind to ponder some of the results. First, no surplus is created on the side of drivers (all drivers earn their reservation wage). Second, the monopsonistic position of a platform in the market for ride-hailing drivers does not result in a markdown in the price paid to drivers.

3.2.3 Matching and wait times

The way platforms match riders and drivers determines how long riders have to wait for their vehicles.[15] I assume platforms match riders to the closest idle driver, which is a natural assumption given that platforms have access to the location of riders and drivers. A driver (or vehicle) is idle if she is not currently busy serving a passenger or en route to pick one up.

Arnott (1996) showed that if I idle vehicles are evenly distributed over a service area of size A, so that the density of idle vehicles is $D = I/A$, the expected distance

[15] Wait time refers to the time it takes a driver to reach the location of the rider she has been matched with. Due to the stochastic nature of the process, there may be some additional wait time if there are no idle vehicles at the time the rider requests a ride. However, for the scale of ride hailing that concerns us (thousands of riders and drivers), this additional wait time is usually only a few seconds, so it is ignored. Li, Tavafoghi, Poolla, and Varaiya (2019) reach a similar conclusion for ride hailing in New York City.

between a rider and the closest idle driver can be approximated as $\frac{1}{2\sqrt{D}}$.[16] Assuming vehicles travel at speed v when en route to pick up a passenger, expected wait time (w) as a function of the density of idle vehicles is:

$$w(D) = \frac{1}{2v\sqrt{D}} \tag{3.5}$$

At any moment, the number of idle vehicles equals the total number of vehicles (d) minus the number of busy vehicles. In steady state, the expected number of busy vehicles equals the product of the number of trips per hour (x) and the average trip time ($t + w$). Mathematically:

$$D = \frac{I}{A} = \frac{d - (t + w)x}{A} \tag{3.6}$$

Plugging this expression into Equation 3.5, expected wait time can be related to the number of riders and drivers according to:

$$w = \frac{1}{2v\sqrt{\frac{d-(t+w)x}{A}}} \tag{3.7}$$

The relationship between expected wait time (w) and the number of riders (x) and drivers (d) given by Equation 3.7 is not simple because w is in both sides of the equation and because in-vehicle time (t) and speed (v) may vary with the number of vehicles due to traffic congestion. In fact, for a given number of riders and drivers, there will usually be two wait times that satisfy Equation 3.7. The highest wait time corresponds to what Castillo et al. (2018) termed "wild goose chases". It describes a situation in which riders are being matched to vehicles that are relatively far, causing vehicles to spend a long time en-route to pick up passengers, which in turn reduces

[16] Appendix C presents a simple proof of this formula. It assumes service area A is big relative to the expected distance to the closest idle vehicle.

the number of idle vehicles and causes the long-distance matches. Although riders could be served by the same number of vehicles at a lower average wait time (the low wait time solution to Equation 3.7), the matching system is stuck at this inefficient equilibrium. Platforms, however, can avoid these "wild goose chases" by limiting the distance of rider-driver matches, a policy adopted by most platforms.[17]

In order to simplify the relationship between wait time, riders and drivers, and to avoid the possibility of "wild goose chases", I assume that total trip time $(t + w)$ in the right-hand side of Equation 3.7 is fixed (independent of the number of riders and drivers).[18] I also assume that the sped of vehicles en-route to pick up passengers is not affected by congestion, which is a plausible assumption considering that these routes usually do not include major roads. Denoting total trip time by s (for service time), Equation 3.7 reduces to:

$$w(x, d) = \frac{1}{2v\sqrt{\frac{d - s \cdot x}{A}}} \tag{3.8}$$

Equation 3.8 now explicitly defines expected wait time as a function of the number of drivers and riders. This equation reveals network effects (or economies of scale) in the ride-hailing industry. If drivers and riders increase in the same proportion, expected wait time declines. For example, a platform with four times as many drivers and riders can offer half wait times in average. The effect is even more transparent when we explicitly take into account that the number of drivers adjust to the number of riders. Using Equation 3.4 to substitute d out of Equation 3.8, we can express

[17] Additionally, the dynamic process that would lead to "wild goose chases" is not clear and the resulting equilibrium is not be stable. The possibility of "wild goose chases" is equivalent to the possibility of hypercongested equilibria in road networks. Small and Verhoef (2007) conclude that this type of equilibria is not adequate to describe road congestion, and that proper analysis of these phenomena requires a dynamic approach (breaking the chains of steady state analysis).

[18] This assumption greatly simplifies the analytical solutions of the model and reveals the economies of scale in the industry. It will not affect the numerical results because I calibrate total trip time to coincide with the sum of in-vehicle and wait times in equilibrium.

expected wait time as a function of only the number of riders:

$$w(x) = \frac{1}{2v\sqrt{\frac{(q/c-s)x}{A}}} \tag{3.9}$$

Equation 3.9 reveals that expected wait time is a decreasing function of the number of riders.

3.2.4 Traffic congestion and in-vehicle time

Additional vehicles on the streets, such as ride-hailing vehicles, tend to slow down all other vehicles and increase travel times. However, to properly judge the effect of ride-hailing vehicles on traffic congestion, we must consider what would happen if ride hailing were not an option for travelers. As explained in Section 4.1, ride hailers may shift to a different mode of transportation or may decide not to travel. If the main alternative for ride hailers is, for example, not to travel or to bike, the effect of ride hailing on traffic congestion would be significant. On the other extreme, if their main alternative is to use private cars, the effect could be small or even negative.[19] The case of public modes such as buses and taxis lies somewhere in between depending on how the supply of public vehicles adjusts to changes in demand.[20]

Expression 3.1 introduced an in-vehicle travel time function ($t(d)$) that depends only on the number of ride-hailing vehicles on the street, which implies all other traffic, such as cars and buses, remains constant as the scale of ride hailing varies. There are two important assumptions behind this simplification. First, there is no substitution between ride hailing and cars (i.e. the outside option for most potential ride hailers is not to use their private cars). Second, the supply of public vehicles that use

[19]The net effect of ride hailing on congestion in comparison to private cars may be negative if private cars have to cruise for parking. On the other hand, ride-hailing vehicles usually deadhead in between trips.

[20]For example, if the number of buses is fixed, a shift of travelers from buses to ride hailing would increase congestion (although it would also benefit bus users due to reductions in crowding and possibly wait times).

common streets (e.g. regular buses and taxis) does not adjust to changes in demand. Section 3.3.1 argues that these assumptions are plausible for Bogotá. However, they may not hold for other cities. Most importantly, if there is significant substitution between ride hailing and cars, the demand model introduced in Section 4.1 should be expanded to explicitly include private car as an option, while in-vehicle time should depend both on the number of ride-hailing vehicles and the number of private cars. These adjustments can have a significant effect on the size of the optimal congestion charge, because demand shifts from ride hailing to private cars are not likely to have a major impact on congestion.[21]

The external cost ride hailing imposes on other road users due to traffic congestion can be approximated as the product of the number of road users, their average value of time and the average in-vehicle time increment caused by ride-hailing vehicles. In turn, this time increment can be approximated as the marginal increase caused by one vehicle times the number of vehicles.[22] Mathematically:

$$EC(d) = \beta_{other} \cdot N_{other} \cdot Mg\_time \cdot d = MEC \cdot d \tag{3.10}$$

where $EC(d)$ is the total external congestion cost caused by ride-hailing vehicles on other road users, N_{other} denotes the average number of other road users (such as private cars or transit passengers) per hour during the time period of analysis, β_{other} represents their average value of time, and Mg_time corresponds to the marginal in-vehicle travel time increase caused by an additional vehicle on the street. $MEC = \beta_{other} \cdot N_{other} \cdot Mg\_time$ compiles the marginal external cost that one ride-hailing vehicle imposes on other road users through traffic congestion. Section 3.3.3 describes the empirical approach I use to estimate Mg_time for the morning peak period in Bogotá.

[21] The optimal congestion charge on ride hailing tends to decrease as substitution with cars intensifies (i.e. as more ride hailers have private car as their best alternative).

[22] Since the relationship between traffic volume and travel time is in general non-linear, this last approximation is valid as long as the number of ride-hailing vehicles represents a small percentage of total traffic.

I will also use this parameter to specify the in-vehicle travel time function for ride hailing ($t(d)$).

It is important to note that Expression 3.10 assumes that the number of other road users, most importantly the number of private cars, remains constant as traffic congestion varies due to changes in the number of ride-hailing vehicles on the streets (zero elasticity). To the extent that the number of private cars responds to changes in congestion, Expression 3.10 overestimates the welfare benefit that a reduction in the number of ride-hailing vehicles causes on other road users, because the initial reduction in congestion is partially offset by an increase in the number of private cars. In an extreme case, the number of private cars adjusts to keep the level of congestion constant (infinite elasticity), so reductions in the number of ride-hailing vehicles do not lead to welfare improvements for other road users.[23] As a result, any potential response by private cars to changes in traffic congestion would tend to decrease the size of the optimal congestion charge.

3.2.5 Equilibrium number of riders and drivers

Equation 3.2 implicitly defines the number of riders or trips per hour (x) as a function of the number of drivers or vehicles (d) and the price per trip charged to riders (p) (demand side). Similarly, Equation 3.4 determines the number of drivers as a function of the number of riders and the price per trip paid to drivers (q) (supply side). These two equations together determine the equilibrium number of riders and drivers (x, d) for any given set of prices (p, q).

For any given set of prices, the point $(x, d) = (0, 0)$ (no riders and no drivers) is always a solution to Equations 3.2 and 3.4. This equilibrium has clear intuition. Without riders, there is no reason for drivers to work. Without drivers, riders would

[23] Even though this extreme case has found some empirical evidence (Duranton & Turner, 2011) and is usually referred to as "the fundamental law of road congestion", it implies very strong assumptions. Specifically, it requires all potential car users to have the same reservation value for driving and the same value of time.

have to be willing to wait forever. However, as long as pricing is sensible (p not too high and/or q not too low), there will be at least one additional equilibrium point with positive numbers of riders and drivers.[24] As long as these additional equilibria exists, I assume the platform is able to reach the equilibrium with the highest number of riders. In practice, platforms would have to take action to avoid the chicken-and-egg dilemma posed by the trivial equilibrium without riders and drivers. For example, platforms can initially guarantee drivers a minimum amount of earnings per hour, which may lead to negative profits in the short run.[25]

3.2.6 Profit- vs Welfare-maximizing prices

Now that we can compute the number of riders and drivers in equilibrium for a given pair of prices, we may proceed to determine the price levels that maximize profit and welfare. Let's start with welfare maximization. Total welfare equals riders' surplus minus the cost of vehicles, drivers and external congestion.[26] The pricing problem for a social planner in control of the platform can then be expressed as:

$$\max_{p,q} \quad \underbrace{N \int_0^{\infty} \int_{p+\beta(t(d)+w(x,d))}^{\infty} [V - \beta(t(d) + w(x,d))] f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta}_{\text{Riders' surplus}} - \underbrace{c \cdot d}_{\substack{\text{Vehicle-driver} \\ \text{cost}}} - \underbrace{MEC \cdot d}_{\substack{\text{External} \\ \text{congestion cost}}} \tag{3.11}$$

where x and d are functions of p and q through the equilibrium process analyzed in the previous section.

Appendix E shows that the welfare-maximizing price gap $(p-q)$ can be expressed

[24] Appendix D presents a simple graphical analysis of these equilibria.

[25] Weyl (2010) proposes insulating tariffs (the price to one side depends on the number of agents on the other side) as a strategy for platforms to avoid coordination failure and implement the desired allocation.

[26] Since the number of riders and other road users are expressed in per-hour units, total welfare and profit are also given on a per-hour basis.

as:

$$p - q = \underbrace{\bar{\beta} \cdot x \cdot \frac{\mathrm{d}(t+w)}{\mathrm{d}x}}_{\substack{\text{Marginal external} \\ \text{cost on ride hailers}}} + \underbrace{\frac{d}{x} \cdot MEC}_{\substack{\text{Marginal external cost} \\ \text{on other road users}}} \tag{3.12}$$

We can interpret Expression 3.12 as a Pigouvian tax. The welfare-maximizing price gap equals the sum of the marginal external cost an additional ride hailer imposes on her fellow ride hailers and on other road users. The marginal cost on ride hailers equals the product of their average value of time ($\bar{\beta}$), their quantity (x) and the marginal effect of an additional ride hailer on total travel time (in-vehicle plus wait, $\mathrm{d}(t+w)/\,\mathrm{d}x$).[27] This cost may turn out to be negative (a benefit) because additional riders increase in-vehicle times through traffic congestion but also reduce wait times due to network effects (see Equation 3.9). The optimal pricing strategy by a social planner may then imply a subsidy and negative revenue ($p - q < 0$), especially if the external cost on other road users is small.[28]

The marginal cost on other road users equals the value of the marginal increase in in-vehicle times caused by ride-hailing vehicles (MEC in Equation 3.10) multiplied by the ratio between vehicles and riders (d/x), which expresses the additional number of vehicles brought by one more rider.

Turning to profit maximization, profit per hour equals the product of the price gap and the number of trips per hour.[29] The pricing problem for a private platform has then a much simpler mathematical form:

$$\max_{p,q} \quad (p - q) \cdot x \tag{3.13}$$

[27] $\mathrm{d}(t+w)/\,\mathrm{d}x$ is *not* a partial derivative. It takes into account the fact that an additional rider causes a proportional increase in the number of ride-hailing vehicles (d).

[28] The potential need for subsidies in public transportation services due to network effects is well known in the transportation literature and is usually referred to as the Mohring effect (Mohring, 1972). See Arnott (1996) for taxis and Parry and Small (2009) for transit.

[29] I assume the costs of developing and maintaining the digital platform are either sunk or independent of the scale of use of the platform (i.e. fixed costs). Consequently, these costs do not affect the profit- or welfare-maximizing price levels.

where x is again a function of p and q through the equilibrium process analyzed in the previous section.

Appendix E shows that the profit-maximizing price gap can be expressed as:

$$p - q = \underbrace{\bar{\beta}_m \cdot x \cdot \frac{\mathrm{d}(t+w)}{\mathrm{d}x}}_{\substack{\text{Marginal external cost on ride hailers valued according to the value of time of marginal riders}}} + \underbrace{\frac{p}{\varepsilon}}_{\text{Markup}} \tag{3.14}$$

Expression 3.14 is again the sum of two terms. The first one is very similar to the first term of the welfare-maximizing price gap, but instead of the average value of time of riders ($\bar{\beta}$), it considers the average value of time of *marginal* riders ($\bar{\beta}_m$).[30] This type of distortion between welfare- and profit-maximizing pricing is not new. Its origins can be traced back to Spence (1975)'s analysis of quality provision by a monopolist. More recently, Weyl (2010) identifies the same distortion, which he names Spence distortion, for multi-sided platforms. It is then not surprising to encounter it in ride-hailing platform pricing.[31] The second term (p/ε) can be interpreted as the usual markup imposed by a monopolist, which is proportional to the inverse of the elasticity of demand.[32]

A comparison of Expressions 3.12 and 3.14 reveals that the difference between the welfare- and profit-maximizing price gaps has three main sources. First, the indifference of a private platform towards the congestion effect on other road users. Second, the tendency of such platform to impose a markup. Third, the Spence distortion in the consideration of the external effect on ride hailers. The first discrepancy tends to make the welfare-maximizing price gap larger, while the second one has the opposite effect. This contest between external effects and markup drives most of the compar-

[30] Marginal riders are those just indifferent between ride hailing and their outside option. They can be found along the ray of Figure 3.1.

[31] The Spence distortion is at the heart of the analysis presented in Chapter 2.

[32] This measure of demand elasticity takes travel time as fixed. See Appendix E for more details.

ison between price gaps. The Spence distortion can go either way, depending on the shape of the bivariate distribution $f(\beta, V)$. In particular, high correlation between reservation values a values of time pushes in favor of a larger welfare-maximizing price gap, because it tends to generate an average value of time of riders ($\bar{\beta}$) larger than that of marginal ones ($\bar{\beta}_m$) (see Section 2.2.5 of Chapter 2).

It is interesting to note that a profit-maximizing platform internalizes (with a Spence distortion) external effects on ride hailers. Currently, ride-hailing vehicles account for only a small portion of vehicle-miles traveled in most cities, so the platform internalizes only a small fraction of the external congestion effect on all road users. However, if ride hailing continues to grow, the platform will internalize a larger fraction of the externality, potentially weakening the motivation for a congestion charge.

3.3 Empirical estimates

This section presents the data and the empirical estimations carried out to calibrate the previous theoretical model to the morning peak period of Bogotá in 2019. I first describe the situation of ride hailing in Bogotá and then detail the two main empirical estimations (demand and marginal congestion).

3.3.1 Ride hailing in Bogotá

Bogotá is a highly dense city of about 7.5 million inhabitants distributed over an urban area of approximately 850km^2. Even though public transportation is the main mode of transportation in the city[33] and a license plate-based restriction takes 50% of the private car fleet out of circulation during peak hours (6-8:30am and 3-7:30pm), Bogotá usually tops worldwide traffic congestion rankings.[34]

[33] Public transportation in Bogotá includes regular buses and a Bus Rapid Transit system (TransMilenio) that operates on exclusive lanes.

[34] Two examples are the Global Traffic Scorecard (Inrix, 2019) and the global analysis based on Google Maps made by P. Akbar, Couture, Duranton, and Storeygard (2020). Both studies rank Bogotá as the most congested city in the world, although Chinese cities were not included.

Uber was the first ride-hailing platform in Bogotá, available since 2013. As in most countries, Uber has faced strong legal challenges in Colombia. The Ministry of Transportation declared ride-hailing services illegal because private vehicles are used to provide a public service, which goes against Colombian law, but the Ministry of Information and Communications Technologies refuses to block the apps based on net neutrality principles. Travelers cannot be penalized for using these services, but drivers can have their driver license suspended temporarily, their car withheld and face monetary penalties. In spite of these difficulties, other platforms such as Beat, Cabify and DiDi followed Uber and are currently available in Bogotá.

The 2019 Mobility Survey of Bogotá provides information about the size of ride hailing in the city for an average weekday (Secretaría Distrital de Movilidad, Bogotá D.C., 2019).[35] Figure 3.2 reveals the number of ride-hailing trips made per hour throughout the day, as well as the percentage this number represents of all trips in the city.[36] Ride hailing peaks in the morning (6-7am) and in the evening (5-6pm) at about 14,000 trips per hour. It accounts for 1 to 2% of all trips most of the day, but its share increases to 15% after midnight (a pattern it shares with taxi trips).

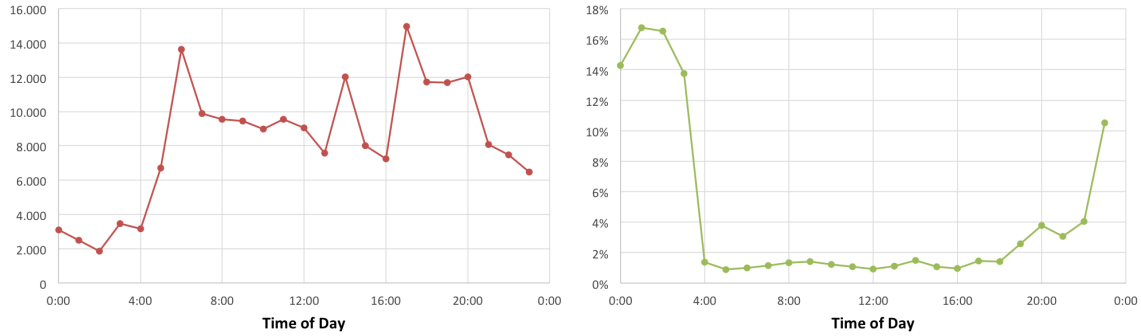
To calibrate the theoretical model developed in Section 3.2, I consider the average size and characteristics of ride hailing during the morning peak period from 6:00 to 8:30am.[37] The average number of ride-hailing trips per hour during this period was

[35]The survey recorded socio-demographic information from a sample of households, as well as detailed information about the trips made by household members the weekday before the data was collected. The sample consists of 21,828 households located in Bogotá and its surrounding municipalities, who were surveyed between February and August. The survey results include weights that make the sample representative of the population. Unless otherwise noted, all the statistics cited in this paper are weighted.

[36]To compute the total number of trips, I exclude walking trips shorter than 15 minutes.

[37]There are advantages and disadvantages to considering shorter (or longer) periods. The main advantage of shorter periods is that average conditions, such as demand and supply levels, are more likely to be representative of the entire period. The main disadvantage is that some of the model assumptions, such as the inability of riders to switch to a different time period, are less likely to hold. I chose a 2.5 hours period to strike a balance between these concerns. Additionally, the 6-8:30am period coincides with the effective period of the license plate-based restriction on private cars, which is important to maintain the model assumption of low substitution between ride hailing and private cars.

Figure 3.2: Ride-hailing trips on an average weekday in Bogotá in 2019.



(a) Trips per hour.

(b) Percentage of all trips.

Note: Trips are classified according to their start times. For example, the number of trips at 6am corresponds to trips that started between 6:00am and 6:59am. The total number of trips considered to compute the percentages excludes walking trips shorter than 15 minutes.

11,300. The average in-vehicle time of these trips was 38.2 minutes; their average wait time, 2.1 minutes; their average distance, 7.66 kilometers; and the average price paid by riders, COL$11,600.[38] The survey also reveals that about 70% of ride-hailing trips during the morning peak used Uber.[39] Importantly, over 95% of ride hailers declared that they did not have a car available for their trip, so turning to their private cars is not their outside option in case ride hailing becomes too expensive. This statistic supports the model assumption of low substitution between ride hailing and private cars, which implies that reductions in the scale of ride hailing lessen traffic congestions. Finally, the survey shows that some areas of the city generate very few ride-hailing trips during the morning peak period. To account for this pattern, I consider a service area of 500km^2 (A in Equation 4.12).

Unfortunately, the survey does not reveal information on the side of drivers, such as the amount paid per trip by platforms or the number of vehicles available for service at different times. However, Azuara, Gonzalez, and Keller (2019) provide an extensive

[38] Using a purchasing power parity (PPP) adjusted conversion rate of $1,340.5 COL/USD (OECD, 2019), the average price translates to about USD$8.6 per trip.

[39] Uber's dominance may have diminished in 2020 due to two events. First, DiDi joined the market in the second half of 2019. Second, Uber had to suspend its services for about a month at the beginning of 2020 due to allegations of illegal competition from taxi unions.

characterization of Uber drivers in Colombia and other Latin American countries. Based on Uber's administrative data, they report that drivers in Colombia generated an average hourly income of COL$14,075 using the platform in January-February 2019.[40] I use this amount as the reservation wage of drivers (c in Equation 3.3).[41] Their study also reveals that drivers use the platform in average 15 hours per week, which suggests that most drivers are available for service only a few hours per day. This statistic supports the model assumption that drivers can choose to work during the morning peak period without having to work on adjacent periods (i.e. driver supply is independent across time periods).

3.3.2 Demand

As explained in Section 4.1, a bivariate distribution $f(\beta, V)$ characterizes the population of potential ride hailers. This distribution represents the heterogeneity in values of time (β) and reservation values (V) across individuals. I assume the distribution has a bivariate normal form, which gives me five parameters to estimate: two means (μ_β, μ_V), two standard deviations (σ_β, σ_V) and one coefficient of correlation (ρ). Population size (N) will be adjusted to achieve, in a monopolistic scenario, the number of ride-hailing trips per hour estimated from the 2019 Mobility Survey.

I estimate these five parameters using data from stated-preference surveys carried out in Bogotá in December 2018 (Oviedo, Granada, & Perez-Jaramillo, 2020). In these surveys, individuals were asked to recall their most recent trip in the city during the morning peak period. They were then asked to choose between the mode of

[40] The study reports an average hourly income of USD$10.5 adjusted by purchasing power parity (PPP). I use a PPP conversion rate of $1,340.5 COL/USD (OECD, 2019) to recover the amount in Colombian pesos. The average was based on the earnings (net of Uber's commission but inclusive of vehicle expenses such as fuel costs) of 1,136 drivers.

[41] The reservation wage of ride-hailing drivers during the morning peak period of Bogotá may differ from this amount due to three concerns. First, even though Bogotá is Uber's largest market in Colombia, the reporte may have included drivers in other Colombian cities. Second, reservation wages are likely to vary throughout the day. Chen et al. (2019) estimate, however, that the reservation wage of Uber drivers in the U.S. during the morning peak period is close to the daily average. Finally, drivers may have different reservation wages for platforms other than Uber.

transportation actually used for the trip and a hypothetical ride-hailing alternative with a specific, and randomly assigned, price (p), in-vehicle time (t) and wait time (w).[42] Let y be a binary variable that takes value 1 if the individual chose ride hailing, 0 otherwise. Each survey observation i can then be summarized by a vector (y_i, p_i, t_i, w_i). After data cleaning, I obtain 1,022 observations for estimation.

The surveys also collected socio-demographic information on each individual. I use these data to compute a weight for each observation (h_i) so that the sample resembles the characteristics of the population of travelers during the morning peak period according to the 2019 Mobility Survey. The characteristics considered to compute the weights include age, gender, socio-economic stratum, transportation mode, trip purpose and trip distance.[43]

The bivariate distribution $f(\beta, V)$ determines travelers' choices between ride hailing and their outside options given a price and in-vehicle travel time for an *average-distance* trip. I then adjust the price and in-vehicle travel time (p_i, t_i) given to each individual in the survey according to their trip distance d_i and the average distance of ride-hailing trips obtained from the 2019 Mobility Survey ($\bar{d} = 7.66km$). For each survey observation, I compute $\tilde{p}_i = (\bar{d}/d_i)p_i$ and $\tilde{t}_i = (\bar{d}/d_i)t_i$.

I estimate the five parameters of the bivariate normal distribution through maximum likelihood. The log-likelihood of the entire sample given a set of parameters $\Theta = (\mu_\beta, \mu_V, \sigma_\beta, \sigma_V, \rho)$ is:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{1,022} h_i \cdot \left[y_i \cdot log \left(\int_0^\infty \int_{\tilde{p}_i + \beta(\tilde{t}_i + w_i)}^\infty f(\beta, V; \Theta)\, \mathrm{d}V\, \mathrm{d}\beta \right) + (1 - y_i) \cdot log \left(1 - \int_0^\infty \int_{\tilde{p}_i + \beta(\tilde{t}_i + w_i)}^\infty f(\beta, V; \Theta)\, \mathrm{d}V\, \mathrm{d}\beta \right) \right] \quad (3.15)$$

[42] If the individual actually made the trip by ride hailing, she was asked if she would continue to do so in case the characteristics for ride hailing were p, t and w instead of those actually experienced during the trip.

[43] I applied an iterative raking procedure to compute the weights.

where $f(\beta, V; \Theta)$ represents the bivariate normal probability density function.

Table 3.1 presents the point estimates and standard errors of the five parameters. The average value of time is COL$9,100/hr, while the average reservation value is COL$12,450.[44] The results also reveal a positive correlation ($\hat{\rho} = 0.63$) between values of time and reservation values. It is reasonable to obtain a positive correlation because income is probably a strong determinant of both values. Individuals with high incomes are expected to have large values of time and reservation values.

Table 3.1: Parameter estimates for the bivariate normal distribution of values of time (β) and reservation values (V).

	μ_β	σ_β	μ_V	σ_V	ρ
Estimate	$9,100/hr	$5,220/hr	$12,450	$5,270	0.63
Std. Err.	$1,160/hr	$1,640/hr	$660	$540	0.15

Notes: All monetary values are in colombian pesos rounded to the nearest ten. PPP adjusted conversion rate $1,340.5 COL/USD. Parameters were estimated through maximum likelihood. Standard errors were computed using the Cramer-Rao bound.

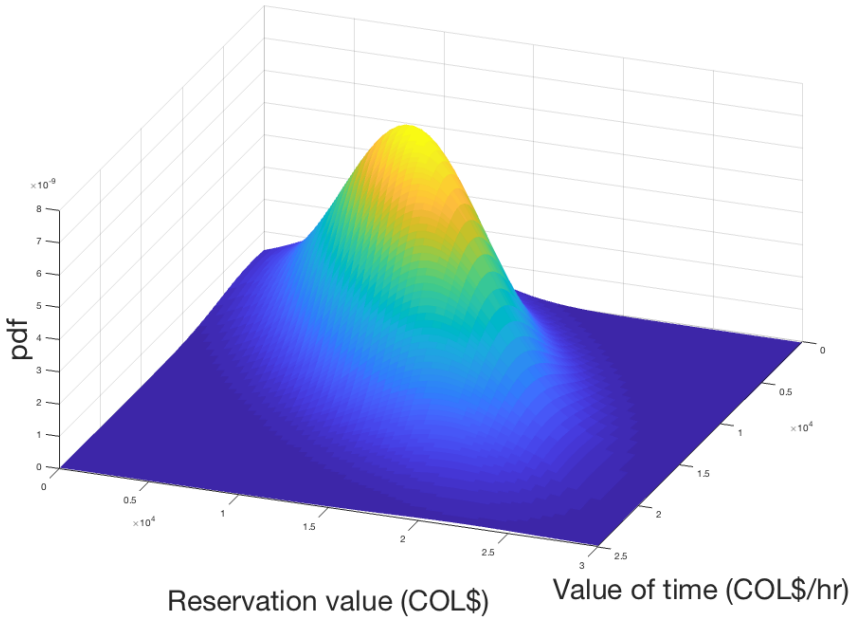
Figure 3.3 graphs the bivariate distribution based on the point estimates. For numerical estimations, I limit the support of the distribution from COL$0 to COL$30,000 in reservation values and from COL$100 to COL$25,000/hr in values of time. I also adjusted the size of the population of potential riders (N) so that the number of trips per hour in the monopolistic scenario reflect the average observed in the 2019 Mobility Survey. The final population size was $N = 145,000$.

3.3.3 Traffic congestion

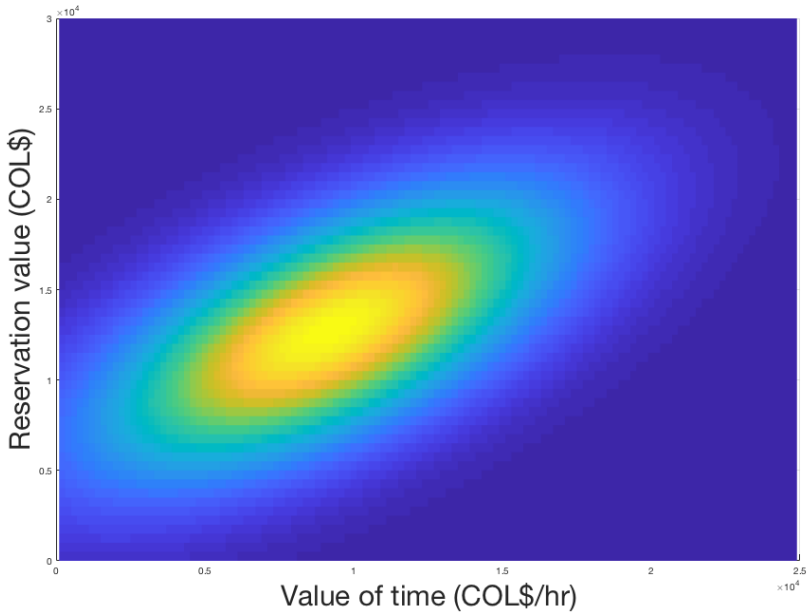
The main objective of the empirical work presented in this section is to identify the marginal effect of additional vehicles on the streets (traffic volume) on average in-vehicle travel times during the morning peak period in Bogotá. In-vehicle travel

[44] Using a purchasing power parity (PPP) adjusted conversion rate of $1,340.5 COL/USD (OECD, 2019), these values translate to an average value of time of USD$6.8/hr and an average reservation value of USD$9.3.

Figure 3.3: Estimated bivariate normal distribution of values of time (β) and reservation values (V).



(a) 3D view.



(b) View from above.

times for a given time period result as an equilibrium outcome of the interaction between travel demand (measured in number of vehicles), which declines as travel times increase, and the capacity of the road network, which dictates how travel times rise as traffic volume grows. To identify the second effect (commonly referred to as the supply side of transportation), we can then use exogenous changes in travel demand, which occur naturally throughout the day as people prefer to (or must) travel at specific times of day.[45]

To illustrate the relationship between travel speed and traffic volume throughout the day, Figure 3.4 displays estimates of both for every 5-min interval of the day based on data from the 2019 Mobility Survey. The figure shows that average speeds oscillate mostly between 20 and 30 km/hr between midnight and 4am, when traffic volume is below 10,000 vehicles, and diminish to about 12 km/hr at peak times, when the number of vehicles rises above 100,000. Traffic volume varies between 70,000 and 100,000 vehicles during most of the day, while average speed stays around 15 km/hr.

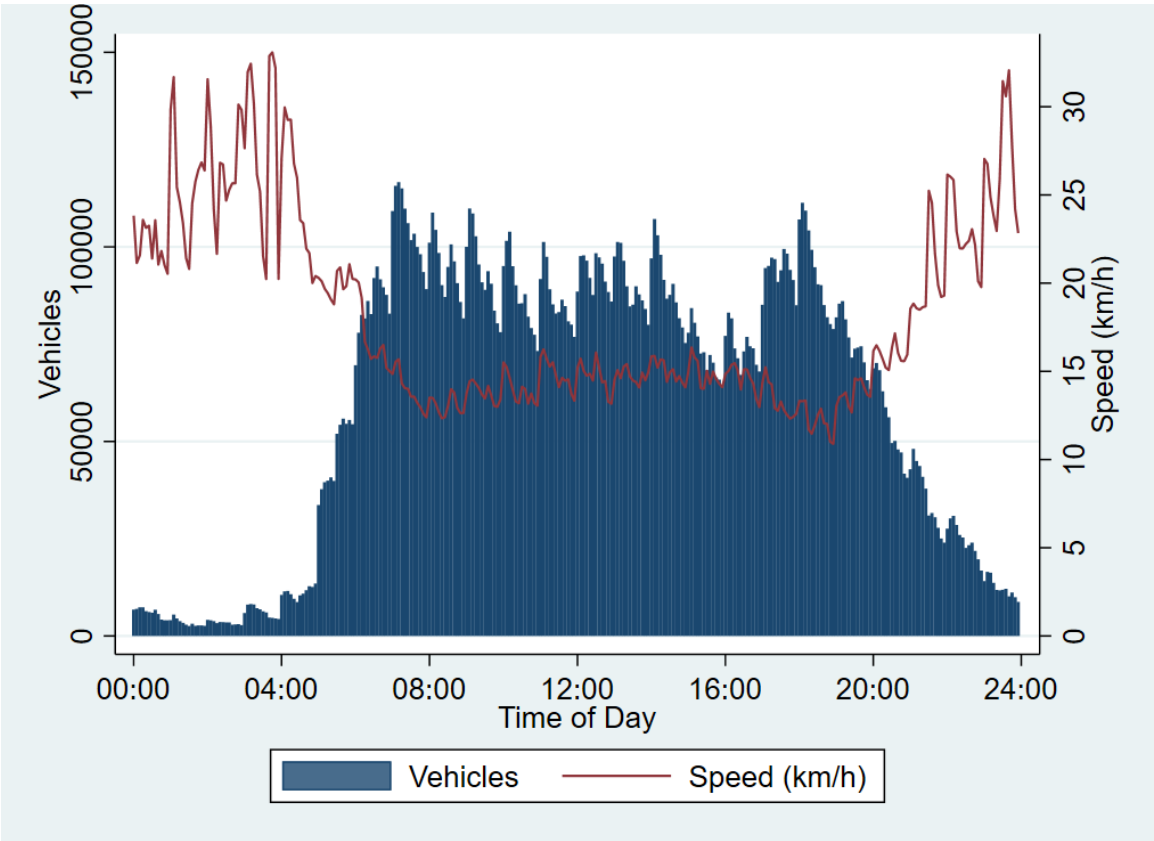
Besides traffic volume, several trip or traveler characteristics may differ across trips taken at different times of day. If these characteristics affect travel speed, they may lead to biases in naive regressions between speed and traffic volume. For example, most trips made during peak periods have mandatory purposes (work or study), and these purposes may encourage travelers to drive faster. To account for these potential effects, I run regressions between speed and traffic volume at the trip level, while controlling for diverse trip and traveler characteristics. The base specification for the regressions is:

$$\frac{1}{speed_i} = \alpha + \beta_v \cdot Veh_i + \vec{\beta}_C \cdot Controls_i + \varepsilon_i \tag{3.16}$$

where $speed_i$ denotes the speed of trip i, Veh_i represents average traffic volume for

[45]The empirical approach employed in this section was originally proposed and applied to Bogotá by P. A. Akbar and Duranton (2017).

Figure 3.4: Traffic volume and average speed in Bogotá (average weekday).



Notes: The number of vehicles includes private cars, taxis and ride-hailing vehicles. The 2019 Mobility Survey differentiates private car trips as driver and as passenger. To compute the number of private cars, I consider only car trips as driver. For taxis and ride-hailing vehicles, I assume one trip corresponds to one vehicle. This assumption may overestimate the number of vehicles to the extent that passengers share rides, but it also may underestimate it to the extent that vehicles deadhead (travel without a passenger). I include trips in a 5-min interval as long they cover any portion of the interval. For example, a trip starting at 8:37am and ending at 9:11am is included in the eight 5-min intervals between 8:35am and 9:15am. I subtract any recorded wait and walk times at origin or destination to consider only in-vehicle travel time. Speed observations come from the same modes. The survey does not report the distance covered by each trip. I approximate these distances by querying Google Maps and obtaining the distance of the recommended route between origin and destination under average traffic conditions. The speed of a 5-min interval corresponds to the average speed of all trips that cover any portion of the interval. Average speeds fluctuate more between midnight and 4am mainly due to less observations for this period.

the duration of trip i,[46] $Controls_i$ include a set of control variables related to the trip (distance, purpose and mode) and the traveler (age, gender and socio-economic stratum), and ε_i is an error term. The dependent variable for the regressions is the inverse of speed, which is proportional to in-vehicle travel time.

The relationship between traffic volume and travel time is usually found to be nonlinear, with travel times increasing more rapidly as traffic volume grows (Small & Verhoef, 2007). To approximate this nonlinear relationship in a flexible manner, I introduce a piecewise linear specification for Veh_i in Equation 3.16. In this specification, the marginal effect of additional vehicles on the inverse of speed (or travel time) may vary for different ranges of traffic volume.

Table 3.2 presents regression results for six different versions of Equation 3.16. Column (6) contains the preferred specification, which includes controls for trip and traveler characteristics and introduces the number of vehicles in a piecewise linear form. The coefficients show that additional vehicles do not affect speed when traffic volume is below 20,000 vehicles, while they have the greatest impact when it rises above 80,000 vehicles. Since I calibrate the theoretical model to the morning peak period, when traffic volumes are above 80,000, the coefficient of interest corresponds to the one for this last range of traffic volume. Its point estimate is 6.98 x 10^{-7} hr/km. This magnitude implies that 10,000 additional vehicles on the streets increase in-vehicle travel time by about 4.2 minutes for a 10 km trip.

As introduced in Equation 3.10, the marginal external cost (MEC) each ride-hailing vehicle imposes on other road users through traffic congestion can be approximated as the product of the number of other road users, their average value of time and the average in-vehicle travel time increment caused by an additional vehicle. I consider two groups of other road users: private cars and taxis.[47] I estimate the

[46] I average the number of vehicles on the streets in all 5-min intervals at least partially covered by the trip.

[47] I ignore the potential impact on public transportation users. TransMilenio uses exclusive bus lanes, while the speed of regular buses depends mainly on the number of stops required to pick up

Table 3.2: Regression results - Inverse of speed on traffic volume.

		(1)	(2)	(3)	(4)	(5)	(6)
Vehicles		5.22***		4.01***		4.29***	
		(0.25)		(0.24)		(0.25)	
Vehicles	<20,000		1.72		0.44		0.08
			(3.80)		(3.59)		(3.58)
	20,000-40,000		5.50*		3.23		4.18
			(2.82)		(2.66)		(2.66)
	40,000-60,000		3.90*		3.61*		3.74*
			(2.29)		(2.16)		(2.15)
	60,000-80,000		6.64***		4.03**		4.04**
			(1.67)		(1.58)		(1.58)
	>80,000		5.64***		6.42***		6.98***
			(1.30)		(1.23)		(1.24)
Controls	Trip	No	No	No	No	Yes	Yes
	Traveler	No	No	Yes	Yes	Yes	Yes

Notes: Each regression estimates Equation 3.16 at the trip level, with the inverse of speed as the dependent variable. All regressions include a constant (α). Columns (1) and (3) include the number of vehicles (Veh_i) as a single variable. Columns (2) and (4) include it in a piecewise linear specification with five ranges. The units of all coefficients are 10^{-7} hr/km. Controls related to the trip include distance and dummy variables that identify the purpose of the trip (mandatory or discretionary) and its mode (car, taxi or ride hailing). Controls related to the traveler include age (<30, 30-50 or >50), gender and socio-economic stratum. Standard errors in parenthesis. Significance levels: $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

number of car and taxi users per hour during the morning peak period based on the 2019 Mobility Survey. For both groups, I use the average value of time estimated in Section 3.3.2 for the population of potential ride hailers (μ_β in Table 3.1). Finally, the average in-vehicle travel time increment caused by an additional vehicle for each group equals the product of their average trip distance (9.3 and 7.5 km for car and taxi trips respectively) and the coefficient obtained from the previous regressions (6.98 x 10^{-7} hr/km).

Similarly, I approximate the in-vehicle travel time function for ride hailing trips (t(d) in Expression 3.1) as the sum of a base time and the increment caused by ride-hailing vehicles through traffic congestion.[48] Finally, I assume ride-hailing vehicles travel at a constant speed of 20 km/hr (not affected by congestion) when en route to pick up a passenger (v in Equation 3.5).[49]

3.4 Results

Table 3.3 reveals the main results of this paper. It presents the optimal pricing decisions of the ride-hailing platform in three scenarios, as well as the outcomes generated by these decisions. In the first scenario, a profit-maximizing firm manages the platform. In the second, it is managed by a social planner, who attempts to maximize overall welfare and so internalizes external congestion effects on other road users. Finally, a private firm takes back control of the platform in the third scenario, but a regulator imposes a tax on the price charged to riders (again with the objective of maximizing overall welfare).[50]

The first scenario resembles the market structure of ride hailing in Bogotá in

and drop off passengers.

[48]The base time represents the expected in-vehicle travel time for a ride-hailing trip of average distance. I adjusted this base time so that in-vehicle time in the monopolistic scenario matches the average observed in the 2019 Mobility Survey for ride-hailing vehicles. The final value was 36.7 minutes. The time increment caused by ride-hailing vehicles equals again the average trip distance (7.66km for ride hailing) times the coefficient obtained from the regressions.

[49]This assumption reflects the notion that vehicles use mainly local uncongested roads when en

Table 3.3: Numerical results for the three main scenarios.

		Monopoly	Social planner	Tax=$1,700
Prices	Riders (p)	$11,980	$13,280	$11,650
	Drivers (q)	$9,770	$9,840	$9,890
Platf. Commission		18.5%	25.9%	15.1%
Quantities	Trips per hour (x)	11,130	5,736	5,707
	Vehicles (d)	7,727	4,011	4,011
Traffic volume increase		7.6%	3.9%	3.9%
In-vehicle time increase		3.9%	2.0%	2.0%
Vehicle utilization		91.5%	89.1%	89.6%
Wait time (min)		2.7	3.7	3.3
Profit (millions/hr)		$24.6	-	$10.0
Public Revenue (millions/hr)		-	$19.7	$9.7
Welfare (millions/hr)	Ride hailing	$45.12	$29.13	$29.07
	Other road users	-$39.84	-$20.68	-$20.68
	Total	$5.27	$8.45	$8.39

Notes: All monetary values are in colombian pesos rounded to the nearest ten. PPP adjusted conversion rate $1,340.5 COL/USD. Prices apply to an average-distance trip (7.66km). The price faced by riders in the last column is COL$13,350 (=$11,650+$1,700).

2019. As mentioned in Section 3.3.1, even though Uber was not the only platform available in Bogotá, it controlled about 70% of the market. The platform charges riders a price per trip of COL$11,980, while it pays drivers COL$9,770 per trip. These values represent prices for an average-distance trip, while the price for specific trips is adjusted in proportion to distance. Since the average distance of a ride-hailing trip during the morning peak period was 7.66 km, the platform's pricing strategy can also be interpreted as charging riders COL$1,564 per km and paying drivers COL$1,275 per km. According to the 2019 Mobility Survey, the average price paid by ride hailers during the morning peak period was about COL$11,600. The monopolistic scenario slightly overestimates this value.[51] The monopolist platform imposes a price gap or platform commission of COL$2,210 or 18.5% (measured as a percentage of the price charged to riders).[52] This gap can be interpreted as the size of the markup imposed by the platform.

As a result of these prices, the number of riders or trips taken per hour is 11,130, while the number of drivers or vehicles available for service is 7,727. Ride-hailing vehicles cause an increase in traffic volume of 7.6%, which raises in-vehicle travel times of all road users by 3.9% in average.[53] The utilization rate of ride-hailing

route to pick up a passenger.

[50]Once a tax τ is imposed by the regulator on the price charged to riders, the private platform still attempts to maximize profit as defined in Equation 3.13, where p represents the price net of tax (i.e. the amount received by the platform). The price effectively paid by riders is $p+\tau$. Accordingly, Equation 3.2, which determines the number of riders, should be adjusted to

$$N \int_{\beta=0}^{\infty} \int_{V=p+\tau+\beta(t(d)+w(x,d))}^{\infty} f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta = x$$

to take into account the effect of the tax on the demand. The regulator then attempts to maximize overall welfare by choosing τ, taking into account how τ affects the profit-maximizing decisions of the private platform.

[51]This difference may be due to the monopolistic market structure imposed on the model, in comparison to a slightly competitive market in reality. But it may also reflect inaccuracies in theoretical assumptions, estimated parameters or Uber's attention to other objectives besides short-run profits (such as market growth).

[52]The commission charged by Uber varies for each trip, but it is thought to be around 20 to 25%. Other platforms, such as DiDi and Cabify, claim to charge lower commissions around 10 to 15%.

[53]Balding, Whinery, Leshner, and Womeldorff (2019) reveal that Uber and Lyft account for 2 to 13% of vehicle-miles traveled (VMT) at the core counties of 6 major U.S. cities. The 7.6% increase in

vehicles, measured as the percentage of time vehicles have a passenger on board, is 91.5%.[54] The average wait time experienced by riders is 2.7 minutes, which is slightly higher than the 2.1 minutes observed from the 2019 Mobility Survey. Finally, the platform gains profits of COL$24.6 millions per hour during the morning peak, while the availability of ride hailing increases overall welfare by COL$5.27 millions per hour. Overall welfare includes net welfare gains from ride hailing ($45.12 millions per hour) minus the external congestion cost imposed on other road users ($39.84 millions per hour).[55] The results suggest that the availability of ride-hailing services in Bogotá during the morning peak period increased overall welfare, in spite of the market suffering from a monopolistic structure and causing traffic congestion externalities.

When a social planner takes control of the platform, her main action is to raise the price charged to riders to COL$13,280, while maintaining the price paid to drivers at about the same level. The price gap then rises to COL$3,440, which constitutes a platform commission of 25.9%. This gap represents mainly the marginal external cost an additional ride hailer imposes on other road users through traffic congestion (second term of Expression 3.12). The marginal external cost on other ride hailers (first term of Expression 3.12) turns out to be slightly negative (about -COL$170) because the benefit of reducing wait times outweighs the cost of increasing in-vehicle times through traffic congestion.

The price increase applied by the social planner has the expected effects. The number of riders and drivers decreases by about 48%. The effect on traffic congestion declines, while average wait times increase. The net welfare gains from ride hailing

traffic volume I find for Bogotá, although indicative, is not entirely comparable to these percentages because I do not include VMT by buses and freight vehicles as part of the total.

[54]Cramer and Krueger (2016) and Balding et al. (2019) report vehicle utilization rates for Uber and Lyft in U.S. cities between 50 and 70%. The utilization rate I find for Bogotá is relatively high in comparison. Two important differences may explain this disparity. First, Bogotá is significantly denser than most U.S. cities, which probably leads to better matching (in terms of distance) between riders and vehicles. Second, I focus on the peak period, while the statistics reported for U.S. cities include off-peak periods and weekends.

[55]Net welfare gains from ride hailing include the welfare created for ride hailers minus the cost of drivers and vehicles (see Equation 3.11).

decrease by about 35%, while the external cost on other road users decreases by 48%. As a result, the overall welfare created by ride hailing is now COL$8.45 millions per hour, 60% more than in the monopolistic scenario. Finally, the platform obtains lower profits (20% reduction), which now constitute public revenue.

A social planner does not have to take control of the platform to realize all the potential welfare gains from ride hailing. In the third scenario, I compute the optimal tax a regulator should impose on the price charged to riders by a profit-maximizing platform in order to maximize overall welfare. The size of the optimal tax or congestion charge is COL$1,700. Again, this value corresponds to the optimal charge for an average-distance trip, while the charge for specific trips should be adjusted in proportion to distance. The optimal charge should be interpreted as COL$222 per kilometer. Alternatively, the optimal charge can be applied as a 14.6% tax on the price charged by the platform to riders. The size of the optimal charge corresponds to 49.4% of the marginal external cost caused by ride hailers, as measured by the optimal price gap when the platform is managed by a social planner.

The size of the optimal charge is larger than the difference between the optimal price gaps in the first two scenarios (COL$1,230). There is a good reason for this discrepancy. The monopolist platform reduces the price it collects from riders as a response to the tax, causing an incomplete pass-through of the tax to riders. In the unregulated scenario, the private platform charges riders COL$11,980 per trip, but when the the regulator imposes the tax the platform decides to collect only COL$11,650 per trip. As a result, the price faced by riders increases from COL$11,980 to COL$13,350 (=$11,650+$1,700) due to the tax. These results imply a monopolistic pass-through of 0.81.[56] To account for this incomplete pass-through, the regulator must impose a tax larger than the initial price gap difference between the profit-maximizing and optimal scenarios.

[56]As a comparison, the empirical study of Leccese (2021) found that a tax on ride hailing in Chicago had a pass-through of more than 100% for single rides and 82% for shared rides.

In this regulated scenario, ride hailing increases overall welfare by COL$8.39 millions per hour, which is very close to the maximum possible (COL$8.45 millions per hour, achieved when a social planner controls the platform). For the regulator to achieve the maximum possible increase in overall welfare, she would have to regulate both sides of the market (e.g. by adding a tax to the price paid to drivers). However, there is little incentive for her to additionally regulate the side of drivers, because the optimal tax on the side of riders already achieves about 98% of the welfare gains available from the unregulated scenario.

Not surprisingly, the tax reduces the profit of the private platform. Profit decreases by almost 60%, while the tax generates public revenue of almost COL$10 millions per hour during the morning peak period.

3.5 Conclusions

By electronically matching riders and drivers, digital platforms raised the efficiency of ride-hailing services and consequently increased their use by urban travelers. Unfortunately, the rapid growth of these services threatens to exacerbate transportation-related externalities in cities around the world, most importantly traffic congestion. Economists have long argued for congestion charges as a tool to mitigate congestion externalities, but their application to private cars has materialized in only a few cities worldwide. However, congestion charges are proving easier to implement for modern ride-hailing services, because platforms already have the technology in place to identify and charge individual trips.

The design of congestion charges for ride-hailing services mediated by digital platforms must consider the structure of these markets. Ride-hailing markets tend to gravitate towards high levels of concentration in very few digital platforms due to positive network effects in wait times. As platforms exert market power to dictate prices, we enter the terrain of environmental regulation under market power. In this

terrain, it is not optimal for private firms to completely internalize external damages, so the size of the optimal Pigouvian tax or congestion charge should be less than the marginal external cost.

In this chapter, I developed and estimated empirically a structural model of ride hailing in order to present the first comparison between market power and congestion externalities for ride-hailing markets. For the morning peak period of Bogotá in 2019, I found that the marginal external cost of congestion is larger than the markup imposed by a monopolist platform. A congestion charge on ride hailing is then justified. However, the optimal size of this charge (as a tax on the price charged by the platform to riders) corresponds to only 50% of the marginal external cost caused by ride hailers. Even though optimal regulation of the ride-hailing industry involves regulating both sides of the market (riders and drivers), the optimal tax on the side of riders achieves 98% of the welfare gains available from the unregulated scenario, so there is little to be gained from additional regulation on the side of drivers.

The comparison between market power and congestion externalities for Bogotá, however, cannot be extrapolated to other cities. It is important to bear in mind that Bogotá is one of the most congested cities in the world. For other less-congested cities, and especially for off-peak periods, the monopolist markup may be larger than the marginal external cost of congestion, in which case a congestion charge on ride hailing would be detrimental. Additionally, the size of the optimal congestion charge is highly sensitive to two important assumptions in the model. First, it was assumed that there is no substitution between ride hailing and private cars (i.e. the outside option for most ride hailers is not to use their private cars). Although this substitution is small for the peak period of Bogotá, probably as a result of the license-plate based restriction in place, it may be significant for other cities with higher car ownership levels. High substitution between ride hailing and private cars reduces the size of the optimal charge. Second, it was assumed that the number of private cars does not

vary as a result of changes in traffic congestion. To the extent that reductions in congestion lead to an increase in the number of private cars, the optimal charge on ride hailing declines, because the benefit from pricing ride-hailing vehicles out of the streets is partially offset by the increase in the number of private cars.

Other results of this chapter are more likely to apply to other cities. Most importantly, I find that an optimal tax on the side of riders achieves most of the available welfare gains, without the need to regulate the supply side of vehicles. As long as there is a very elastic supply side (i.e. the number of drivers adjusts to maintain a constant level of earnings per hour), this finding should be relevant to other cities. Additionally, I find that a monopolist platform passes through about 80% of the tax to riders, which implies that the tax is effective at reducing the scale of ride hailing.

Chapter 4

Modeling Competition Between Ride-Hailing Platforms

4.1 Introduction

The existence of network effects in wait times suggests that ride-hailing markets will gravitate towards a single platform. In fact, many cities around the world already experience a monopolistic ride-hailing market. For example, Didi controls the entire market in Chinese cities, Grab does the same in many cities of southeast Asia, while Uber dominates some of the largest European cities. However, not all cities seem destined to be controlled by a single platform. The most relevant example is that of U.S. cities. Even though Uber's market share in U.S. cities grew steadily until 2016 (Smichowski, 2018), it has remained relatively constant around 70% since 2017, while Lyft controls the remaining 30% (Statista, 2020), which suggests a duopolistic market structure.[1]

This chapter analyzes the characteristics of a ride-hailing market under a duopolistic structure. I first show that without differentiation between platforms (beyond the potential endogenous differentiation created by wait times), duopoly competition leads to zero profits. This result supports the idea that ride-hailing markets will gravitate towards a single platform absent any differentiation between platforms. I then add a small amount of differentiation between platforms. The duopoly equi-

[1] Uber's market share stopped increasing at the same time the company faced public scrutiny related to its organizational culture, which suggests Uber acquired a negative public image.

librium reached by differentiated platforms reduces the price charged to riders and significantly increases the size of the market in comparison to the monopolistic scenario. The profit earned by each platform declines and, even though the market is more competitive, overall welfare also declines due to the external effect on traffic congestion. The optimal congestion charge for a duopoly is then 34% larger than the optimal charge for a monopolistic platform.

This chapter also explores the effect on platform competition of modifying two assumption of the ride-hailing model in Chapter 3. First, I adjust the model to reflect the fact that modern ride-hailing platforms allow travelers to check the location of the closest idle vehicle before deciding to hail a ride. Travelers can then base their decisions on realized instead of expected wait time. Even though this adjustment does not lead to significant changes in the prices set by platforms, it does improve the matching efficiency (lower average wait times), which allows private platforms to earn higher profits and regulators to achieve higher welfare gains.

Second, I change the decision variable for platforms from price to quantity on the supply side of the market (i.e. platforms choose directly the number of vehicles to have available for service). This adjustment is motivated by the advent of autonomous vehicles. When platforms choose directly the number of vehicles to have available for service, the market becomes less competitive. The price charged to riders increases, the size of the market decreases and profits rise. However, the impact on overall welfare is positive, because the external effect on traffic congestion diminishes.

The rest of this chapter is organized in two main sections. The first section introduces the use of realized wait times in the context of a single platform. The second section presents the analysis of a duopolistic structure, including the effect on platform competition of moving to realized wait times and quantity competition on the side of vehicles. The last section presents the main conclusions.

4.2 Expected vs realized wait times

This section introduces a change in the ride-hailing model developed in Chapter 3 that aims to more accurately reflect the experience travelers have with modern ride-hailing platforms. In the model of Chapter 3, travelers decide to hail rides based on the average or *expected* wait time to the closest idle vehicle. This approach was initially developed by Arnott (1996) for radio-dispatched taxi systems and has been widely applied to modern ride-hailing platforms (see for instance Castillo et al. (2018) and Li et al. (2019)). However, there is an important difference between radio-dispatched taxi systems and modern ride-hailing platforms in terms of the information available to travelers at the time they decide to hail a ride. With modern platforms, travelers can check the location of the closest idle vehicle on their smartphones before deciding to hail a ride, so they can base their decisions on *realized* instead of expected wait times.

In the following subsections, I first summarize the model developed in Chapter 3. I then introduce the necessary adjustments to the model to reflect travelers basing their decisions to hail rides on realized wait times. The last subsection describes how these adjustments affect the profit- and welfare-maximizing prices and outcomes obtained in Chapter 3 for the morning peak period of Bogotá, which assumed that a single platform controlled the entire market. Section 4.3.2 will present how the move from expected to realized wait times affects competition between platforms.

4.2.1 Summary of the ride-hailing model based on expected wait times

The ride-hailing model developed in Chapter 3 had four main components. On the demand side, the population of potential ride hailers was characterized by a bivariate distribution of reservation values and values of time $f(\beta, V)$. Given an in-vehicle travel time t, an average or expected wait time w and a price per trip p, traveler i

decides to hail a ride only if $V_i - \beta_i(t+w) - p \geq 0$. The number of ride hailers x can then be computed as

$$x = N \int_{\beta=0}^{\infty} \int_{V=p+\beta(t+w)}^{\infty} f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta \tag{4.1}$$

where N is the size of the population of potential ride hailers.

On the supply side, the number of drivers or vehicles adjusts to maintain a fixed level of hourly earnings c (the reservation wage of drivers). Given a price per trip q paid by the platform to drivers, the number of drivers d can be computed as

$$d = \frac{q}{c} \cdot x \tag{4.2}$$

Average or expected wait time w is a function of the density of idle vehicles D, which in turn depends on the total number of drivers d and riders x in the market according to the following equations:

$$w = \frac{1}{2v\sqrt{D}} = \frac{1}{2v\sqrt{\frac{d - s \cdot x}{A}}} \tag{4.3}$$

where v is the speed of idle vehicles when en route to pick up a passenger, s is the average service time of each trip (in-vehicle plus wait) and A is the size of the service area for the market.

Finally, in-vehicle time t is an increasing function of the number of ride-hailing vehicles d due to traffic congestion. I specify this function linearly, considering that ride-hailing vehicles are usually a small portion of total traffic volume. In-vehicle time can then be expressed as

$$t = b + mg\_time * d \tag{4.4}$$

where b is the base in-vehicle time for ride hailing (achieved when the number of ride-

hailing vehicles approaches zero) and mg_time is the marginal increase in in-vehicle time caused by an additional vehicle on the roads due to traffic congestion.

Equations 4.1 to 4.4 determine the number of riders and drivers in the market in equilibrium, as well as in-vehicle and wait times, for a given set of prices (p, q) charged to riders a paid to drivers per trip by the platform. As analyzed in Appendix D, a solution without riders nor drivers ($x = 0$, $d = 0$) is always a potential equilibrium. However, as long as pricing is sensible (p not too high and/or q not too low) there will be at least one additional equilibrium point with positive numbers of riders and drivers, which I assume is the one reached by the platform.

A private firm in charge of the platform chooses prices to maximize profit

$$\max_{p,q} \quad (p - q) \cdot x \tag{4.5}$$

A social planner in charge of the platform chooses prices to maximize overall welfare, which includes the net welfare created by ride hailing minus the external cost imposed on other road users through traffic congestion

$$\max_{p,q} \quad \underbrace{N \int_0^{\infty} \int_{p+\beta(t+w)}^{\infty} [V - \beta(t+w)] f(\beta, V) \, \mathrm{d}V \, \mathrm{d}\beta}_{\text{Riders' surplus}} - \underbrace{c \cdot d}_{\substack{\text{Vehicle-driver} \\ \text{cost}}} - \underbrace{MEC \cdot d}_{\substack{\text{External} \\ \text{congestion cost}}} \tag{4.6}$$

where MEC is the marginal external cost an additional ride-hailing vehicle imposes on other road users.

4.2.2 Ride-hailing model based on realized wait times

The previous model assumes travelers base their decisions to hail rides on the average or expected wait time to the closest idle vehicle. However, modern ride-hailing platforms usually allow potential riders to check the location of the closest idle vehicle and give them a wait time estimate before they decide to hail a ride. Travelers can then base their decisions on realized wait times. For example, even if there are a lot of idle vehicles (leading to a low expected wait time), an unlucky traveler may be located far from the closest one. Her realized wait time would then be long and she would probably decide to go for her outside option (which may be not to travel or to use other mode of transportation).

In order to introduce this behavioral change into the model, we must first realize that a density of idle vehicles D determines not only an average wait time (given by Equation 4.3) but an entire probability distribution. Appendix C shows that given a density of idle vehicles D, the probability that the closest idle vehicle to a rider is located at a distance shorter than y is

$$pr(r \leq y) = 1 - e^{-\pi y^2 D} \tag{4.7}$$

where r is the random variable that represent distance to the closest idle vehicle (with support $[0, \infty)$). Considering that vehicles travel at speed v when en route to pick up a passenger, the probability that wait time w is below a given time threshold h is

$$pr(w \leq h) = 1 - e^{-\pi v^2 h^2 D} \tag{4.8}$$

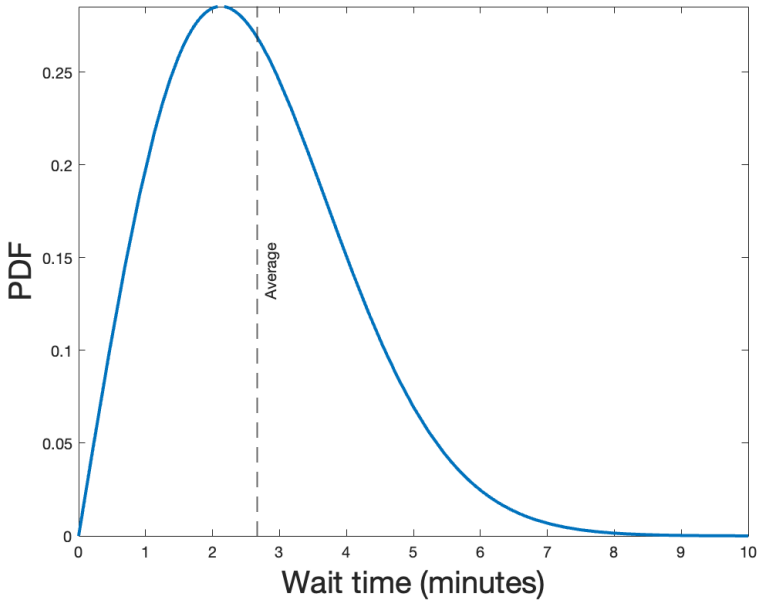
which implies (by differentiation of the previous cumulative density function) that

the probability density function for wait time is

$$f(w) = 2\pi v^2 w D e^{-\pi v^2 w^2 D} \tag{4.9}$$

To visualize the previous distribution, Figure 4.1 graphs the probability density function of wait time for a density of idle vehicles $D = 0.3172 km^{-2}$.[2] At this density, the average wait time is 2.66 minutes, but realized wait time varies mostly between 0 and 8 minutes.

Figure 4.1: Probability density function of wait time.



Note: The density of idle vehicles is $D = 0.3172 km^{-2}$.

A traveler with reservation value V and value of time β decides to hail a ride only if $V - \beta(t + w) - p \geq 0$, which implies that the maximum wait time she is willing to accept in order to choose ride hailing is $(V - \beta t - p)/\beta$. The closest idle vehicle must then be at a distance shorter than $(V - \beta t - p)v/\beta$ for her to hail a ride. Denoting this maximum distance by y, the probability that she chooses to hail a ride

[2] This is the density of idle vehicles achieved in equilibrium in the unregulated scenario of Chapter 3 (first column of Table 3.3).

is $1 - e^{-\pi y^2 D}$ (from Equation 4.7). Given a density of idle vehicles D, the expected number of riders can then be computed as

$$x = N \int_{\beta=0}^{\infty} \int_{V=p+\beta t}^{\infty} \left(1 - e^{-\pi y^2 D}\right) f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta \tag{4.10}$$

where

$$y = \frac{(V - \beta t - p)v}{\beta} \tag{4.11}$$

Recall that the density of idle vehicles is a function of the total number of ride-hailing vehicles and the number of riders according to

$$D = \frac{d - sx}{A} \tag{4.12}$$

When travelers base their decisions to choose ride hailing on realized wait times, Equations 4.10 and 4.12 replace Equations 4.1 and 4.3 to define, together with Equations 4.2 and 4.4, the expected number of riders and drivers in equilibrium for a given set of prices (p, q).

The profit-maximization problem for a private platform can still be stated as in Expression 4.5, which in this case represents expected profits. To compute expected overall welfare, however, we must first compute the expected wait time that each traveler faces *conditional* on choosing to hail a ride. Appendix F shows that a traveler who decides to hail a ride only if the closest available vehicle is at a distance shorter than y faces an expected wait time given by

$$E[w|y] = \frac{\frac{\mathrm{erf}\left(\sqrt{\pi D} y\right)}{2\sqrt{D}} - y e^{-\pi y^2 D}}{v\left(1 - e^{-\pi y^2 D}\right)} \tag{4.13}$$

where $\text{erf}(\cdot)$ is the error function.[3]

The expected welfare-maximization problem for a social planner in charge of the platform can then be expressed as

$$\max_{p,q} \quad \underbrace{N \int_0^\infty \int_{p+\beta t}^\infty \left(1 - e^{-\pi y^2 D}\right) [V - \beta(t + E[w|y])] f(\beta, V) \, \mathrm{d}V \, \mathrm{d}\beta}_{\text{Riders' surplus}} - \underbrace{c \cdot d}_{\substack{\text{Vehicle-driver} \\ \text{cost}}} - \underbrace{MEC \cdot d}_{\substack{\text{External} \\ \text{congestion cost}}} \tag{4.14}$$

4.2.3 Comparison of results for a single platform

This section explores numerically the differences on the profit- and welfare-maximizing prices and outcomes imposed by a single platform in control of an entire ride-hailing market caused by having travelers base their decisions to hail rides on expected or realized wait times. I use the parameter estimates from Chapter 3, which were calibrated to the morning peak period of Bogotá in 2019 (see Section 3.3).

Table 4.1 compares the results when the platform is managed by a private firm (profit maximization) and when the platform is managed by a social planner (welfare maximization). In both cases, the move from expected to realized wait times decreases slightly (by less than 1%) the prices charged to riders and paid to drivers by the platform. As a result, the expected number of riders and drivers increases slightly. Even though the higher number of riders and drivers implies a higher unconditional expected wait time for riders (Equation 4.3), the fact that travelers condition their decision to hail a ride on realized wait times leads to lower average wait times in

[3] The error function is defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, \mathrm{d}t$$

both scenarios.[4] This effect is more significant in the welfare-maximization scenario because wait times tend to be larger in this scenario due to the smaller scale of the market.

Table 4.1: Comparison of results based on expected and realized wait times for a single platform.

		Profit Max.		Welfare Max.	
		Expected	Realized	Expected	Realized
Prices	Riders (p)	\$11,984	\$11,970	\$13,281	13,251
	Drivers (q)	\$9,772	\$9,741	\$9,841	9,804
Platf. Commission		18.46%	18.62%	25.90%	26.01%
Quantities	Trips per hour (x)	11,130	11,220	5,736	5,848
	Vehicles (d)	7,727	7,765	4,011	4,073
In-vehicle time increase		3.87%	3.89%	2.01%	2.04%
Average wait time (min)		2.66	2.56	3.71	3.52
Profit (millions/hr)		\$24.63	\$25.01	\$19.73	\$20.16
Overall welfare (millions/hr)		\$5.27	\$5.61	\$8.45	\$8.80

Notes: All monetary values are in colombian pesos. PPP adjusted conversion rate \$1,340.5 COL/USD. Prices apply to an average-distance trip (7.66km).

The move from expected to realized wait times allows the private firm to increase profit by 1.5%, and the social planner to increase the overall welfare created by ride hailing by 4.1%.

4.3 Duopoly

So far, I have always assumed that a single platform controls the ride-hailing market. This monopolistic assumption reflects the situation of ride-hailing in many cities around the world (e.g. Grab in southeast Asia and DiDi in China). Additionally, the existence of economies of scale or network effects in wait times provides a strong

[4]The unconditional expected wait time for riders when travelers consider realized wait times is 2.70 minutes in the profit-maximization scenario and 3.82 minutes in the welfare-maximization scenario.

reason why ride-hailing markets gravitate towards a single platform. However, more than one platform may coexist if they manage to differentiate their services beyond the potential endogenous differentiation caused by wait times (i.e. if some riders have a preference for one platform and some for another when their prices and wait times are the same).

In this section, I extend the ride-hailing models of Section 4.2 to consider two competing platforms. In the first subsection, I show that if there is no inherent differentiation between the platforms (beyond the potential endogenous differentiation created by wait times), competition leads to zero profits for both platforms. Considering that platforms incur in non-negligible fixed costs, this result supports the notion that ride-hailing markets gravitate towards a single platform. I then introduce varying degrees of differentiation between platforms and explore how they affect the resulting equilibrium prices and outcomes, as well as the size of the optimal congestion charge. Finally, I compare the equilibrium results of this differentiated duopoly setting when riders base their decisions on expected wait times against the results when they consider realized wait times.

4.3.1 No differentiation between platforms

When two ride-hailing platforms are available, travelers can choose either one of them or their outside option (other mode of transportation or not traveling). Normalizing the value of the outside option to zero, the values traveler i gets from each option are

$$\text{Platform } 1 : V_i - \beta_i \cdot [t + w_1] - p_1$$

$$\text{Platform } 2 : V_i - \beta_i \cdot [t + w_2] - p_2$$

$$\text{Outside option} : 0$$

where V_i is the reservation value traveler i assigns to ride hailing, which is the same for both platforms due to lack of differentiation; β_i is her value of time, also assumed equal for the two platforms; t is in-vehicle travel time, which is the same for the two platforms because traffic congestion affects them in the same manner; w_1 and w_2 are the wait times offered by each platform; and p_1 and p_2 are the prices charged by the platforms to riders per trip.[5] As usual, the population of potential riders is characterized by a bivariate distribution of reservation values and values of time $f(\beta, V)$.

Assuming drivers have the same reservation wage c working for either platform, the number of drivers d_1 and d_2 for each platform is given by

$$d_1 = \frac{q_1}{c} \cdot x_1; \qquad d_2 = \frac{q_2}{c} \cdot x_2 \tag{4.15}$$

where q_1 and q_2 are the prices per trip paid by each platform to drivers; and x_1 and x_2 are the number of riders (per unit of time) on each platform.

Since the vehicles of both platforms contribute equally to traffic congestion, in-vehicle travel time is now a function of the total number of vehicles

$$t = b + mg\_time * (d_1 + d_2) \tag{4.16}$$

where b and mg_time denote again the base in-vehicle time and the marginal increase caused by one vehicle.

As introduced in Section 4.2, travelers may consider expected or realized wait times when choosing among the two platforms and their outside option. These two behavioral assumptions lead again to two different ways to compute the expected number of riders on each platform.

[5] As usual, these prices should be interpreted as prices for an average-distance trip, while specific prices are proportionally adjusted to distance.

Expected wait times

The expected wait time for each platform is again a function of the number of riders and drivers on each one

$$w_1 = \frac{1}{2v\sqrt{\frac{d_1 - s_1 \cdot x_1}{A}}}; \qquad w_2 = \frac{1}{2v\sqrt{\frac{d_2 - s_2 \cdot x_2}{A}}} \tag{4.17}$$

where s_1 and s_2 are the average service times on each platform (in-vehicle plus wait); v is the speed of vehicles when en route to pick up a passenger; and A is the size of the service area.

Assuming that Platform 2 charges a higher price per trip and offers a lower expected wait time, Figure 4.2 characterizes the distribution of riders between the two platforms.[6] Riders self-select into the two platforms based on their values of time. Riders with high values of time (above threshold β^*) prefer Platform 2, while riders with low values of time prefer Platform 1.[7]

The number of riders on each platform can then be computed as

$$x_1 = N \int_0^{\beta^*} \int_{p_1 + \beta(t+w_1)}^{\infty} f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta \tag{4.18}$$

$$x_2 = N \int_{\beta^*}^{\infty} \int_{p_2 + \beta(t+w_2)}^{\infty} f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta \tag{4.19}$$

where β^* identifies the value of time threshold that separates the riders on each platform, which can be computed from the following equality

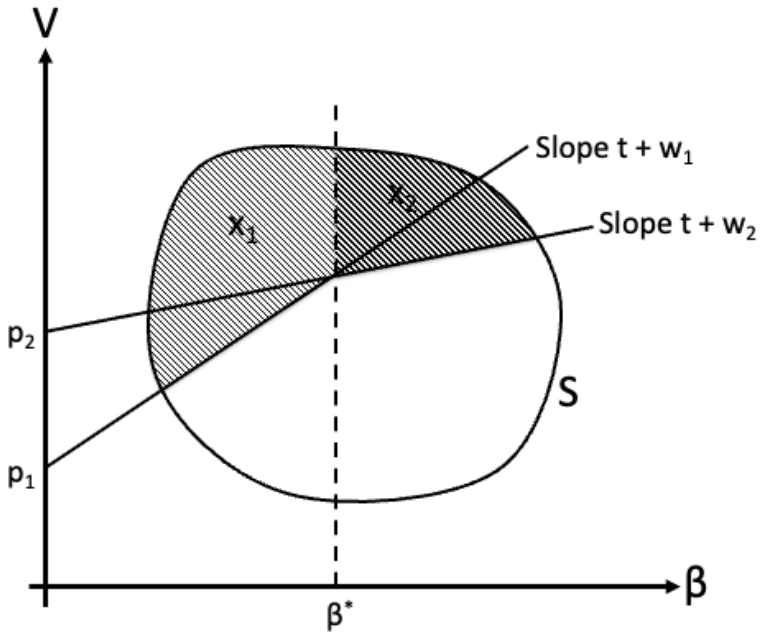
$$p_1 + \beta^* w_1 = p_2 + \beta^* w_2 \tag{4.20}$$

Equations 4.18 and 4.19 determine, together with Equations 4.15 to 4.17, the

[6] If the platform with the higher price also offers a higher expected wait time, it would clearly attract no riders.

[7] This self-selection pattern was introduced in Section 2.2.7 for general congestible resources.

Figure 4.2: Distribution of riders between two platforms based on expected wait times.



Notes: S represents the support of the bivariate distribution of values of time (β) and reservation values (V) across potential riders. x_1 and x_2 identify the number of riders on each platform. p_1, p_2 and w_1, w_2 denote the prices per trip charged and expected wait times offered by each platform. β^* represents the value of time threshold that separates the riders on each platform.

number of riders (x_1, x_2) and drivers (d_1, d_2) on each platform in equilibrium (as well as expected wait times and in-vehicle time) for a given set of prices (p_1, q_1) and (p_2, q_2) set by the platforms. As usual, these equations may offer multiple equilibria because solutions involving no riders and no drivers on one or both platforms are always possible. However, as long as equilibria with positive numbers of riders and drivers for one or both platforms exist, I assume platforms reach these equilibria.

Realized wait times

The density of idle vehicles on each platform can be computed as

$$D_1 = \frac{d_1 - s_1 x_1}{A}; \qquad D_2 = \frac{d_2 - s_2 x_2}{A} \tag{4.21}$$

As introduced in Section 4.2.2, the density of idle vehicles on a platform determines a probability distribution of wait times for potential riders (with support $[0, \infty)$). Travelers can then check the realized wait time on each platform before deciding which platform (if any) to use. Assuming $p_2 > p_1$, travelers for whom $V - \beta t - p_1 < 0$ will surely not use either platform. Travelers for whom $V - \beta t - p_1 > 0$ but $V - \beta t - p_2 < 0$ will surely not use Platform 2, but may use platform 1 if its realized wait time is low enough. Finally, travelers for whom $V - \beta t - p_2 > 0$ may use either platform depending on their realized wait times. The expected number of riders on each platform can then be computed as

$$\begin{aligned} x_1 = N \int_0^{\infty} \int_{p_1+\beta t}^{p_2+\beta t} pr(\beta, V, p_1, D_1) \cdot f(\beta, V) \, \mathrm{d}V \, \mathrm{d}\beta \\ + N \int_0^{\infty} \int_{p_2+\beta t}^{\infty} pr_1(\beta, V, p_1, p_2, D_1, D_2) \cdot f(\beta, V) \, \mathrm{d}V \, \mathrm{d}\beta \end{aligned} \tag{4.22}$$

$$x_2 = N \int_0^{\infty} \int_{p_2+\beta t}^{\infty} pr_2(\beta, V, p_1, p_2, D_1, D_2) \cdot f(\beta, V) \, \mathrm{d}V \, \mathrm{d}\beta \tag{4.23}$$

where $pr(\beta, V, p_1, D_1)$ is the probability that a traveler with reservation value V and

value of time β chooses Platform 1, given that Platform 2 is not an option due to its price. This probability is a function of the price and density of idle vehicles of Platform 1, as well as the individual's reservation value and value of time. From Section 4.2.2, we know this probability can be expressed as

$$pr(\beta, V, p_1, D_1) = 1 - e^{-\pi y_1^2 D_1} \tag{4.24}$$

where $y_1 = \frac{(V-\beta t-p_1)v}{\beta}$ represents the maximum distance to the closest idle vehicle that the traveler is willing to accept to use Platform 1.

$pr_1(\beta, V, p_1, p_2, D_1, D_2)$ and $pr_2(\beta, V, p_1, p_2, D_1, D_2)$ denote the probabilities that a traveler chooses each platform, given that either platform may be chosen depending on their realized wait times. Besides the traveler's reservation value and value of time, these probabilities depend on the prices and densities of idle vehicles of the platforms. The analytical expressions for these probabilities are significantly more complicated than Expression 4.24. Appendix G derives these analytical expressions.

Note that even if the platform charging the highest price offers a lower density of idle vehicles, and so longer wait times in average, it may still attract riders due to the probabilistic nature of the assignment. A traveler may choose this platform if one of its idle vehicles happens to be very close to its location. If travelers base their decisions on expected wait times, such a platform would not attract riders.

Equations 4.22 and 4.23 determine, together with Equations 4.15, 4.16 and 4.21, the expected number of riders (x_1, x_2) and drivers (d_1, d_2) on each platform in equilibrium (as well as idle-vehicle densities and in-vehicle time) for a given set of prices (p_1, q_1) and (p_2, q_2) set by the platforms.

Duopoly equilibrium

Appendix H shows that, in the previous duopoly settings, if one of the platforms sets prices (p, q) and the other $(p, q + \epsilon)$, where ϵ is a positive but potentially very small amount, only the platform with the highest price paid to drivers per trip can obtain a positive number of riders and drivers in equilibrium (i.e. it wins the entire market).[8] This result mirrors the usual duopoly setting of Bertrand competition in prices. As a result, none of the platforms can achieve positive profit in equilibrium. To the extent that platforms face fixed costs of operation (e.g. administrative costs), this perfect-competition result suggests that ride-hailing markets without inherent differentiation across platforms will gravitate towards a single platform.[9]

4.3.2 Differentiated platforms

Beyond price and wait time, other characteristics of ride-hailing trips may cause travelers to prefer one platform over another. For example, travelers may regard one platform as safer or more reliable. These travelers may then choose this platform even if it offers an inferior price-wait time combination. Assuming that these additional characteristics are independent of the scale of the platform (i.e. exogenous), their implication in terms of the previous duopoly settings is that travelers could have different reservation values for the two platforms.

Denoting the reservation values of traveler i for each platform as V_{1i} and V_{2i}, the

[8] The number of riders and drivers on the platform with the highest price paid to drivers will be positive as long as the price pair $(p, q + \epsilon)$ is sensible (p not too high and/or q not too small, see Appendix D).

[9] Formally, consider a two-stage game between two ride-hailing platforms. In stage 1, platforms decide to enter the market or not. If a platform enters the market, it incurs a positive fixed cost. In stage 2, platforms compete in prices (p, q) (in case both decide to enter), or a single platform sets monopolistic pricing. Assuming that monopolistic profit is higher than the fixed cost, the only equilibrium of this game has only one platform in the market (Mas-Colell, Whinston, & Green, 1995, Section 12.E).

values traveler i gets from each option are

$$\text{Platform 1} : V_{1i} - \beta_i \cdot [t + w_1] - p_1$$

$$\text{Platform 2} : V_{2i} - \beta_i \cdot [t + w_2] - p_2$$

$$\text{Outside option} : 0$$

The population of potential riders is now characterized by a trivariate distribution of reservation values and values of time $f(\beta, V_1, V_2)$. The case of undifferentiated platforms analyzed in the previous section can now be regarded as a special case in which V_1 and V_2 are perfectly correlated across the population of potential riders.

Equations 4.18 and 4.19, which determine the number of riders on each platform as a function of their prices and expected wait times, must now be extended to

$$x_1 = N \int_0^\infty \int_{p_1+\beta(t+w_1)}^\infty \int_0^{V_1+(p_2-p_1)+\beta(w_2-w_1)} f(\beta, V_1, V_2)\, \mathrm{d}V_2\, \mathrm{d}V_1\, \mathrm{d}\beta \tag{4.25}$$

$$x_2 = N \int_0^\infty \int_{p_2+\beta(t+w_2)}^\infty \int_0^{V_2-(p_2-p_1)-\beta(w_2-w_1)} f(\beta, V_1, V_2)\, \mathrm{d}V_1\, \mathrm{d}V_2\, \mathrm{d}\beta \tag{4.26}$$

Equations 4.25 and 4.26 determine, together with Equations 4.15 to 4.17, the number of riders (x_1, x_2) and drivers (d_1, d_2) on each platform in equilibrium (as well as expected wait times and in-vehicle time) for a given set of prices (p_1, q_1) and (p_2, q_2) set by the platforms when travelers base their decision on expected wait times.

When travelers base their decision on realized wait times, Equations 4.22 and 4.23 must be extended to

$$\begin{aligned} x_1 = {} & N \int_0^\infty \int_{p_1+\beta t}^\infty \int_0^{p_2+\beta t} pr(\beta, V_1, p_1, D_1) \cdot f(\beta, V_1, V_2)\, \mathrm{d}V_2\, \mathrm{d}V_1\, \mathrm{d}\beta \\ & + N \int_0^\infty \int_{p_1+\beta t}^\infty \int_{p_2+\beta t}^\infty pr_1(\beta, V_1, V_2, , p_1, p_2, D_1, D_2) \cdot f(\beta, V_1, V_2)\, \mathrm{d}V_2\, \mathrm{d}V_1\, \mathrm{d}\beta \end{aligned} \tag{4.27}$$

$$x_2 = N \int_0^{\infty} \int_{p_2+\beta t}^{\infty} \int_0^{p_1+\beta t} pr(\beta, V_2, p_2, D_2) \cdot f(\beta, V_1, V_2)\, \mathrm{d}V_1\, \mathrm{d}V_2\, \mathrm{d}\beta$$
$$+ N \int_0^{\infty} \int_{p_2+\beta t}^{\infty} \int_{p_1+\beta t}^{\infty} pr_2(\beta, V_1, V_2, , p_1, p_2, D_1, D_2) \cdot f(\beta, V_1, V_2)\, \mathrm{d}V_1\, \mathrm{d}V_2\, \mathrm{d}\beta \quad (4.28)$$

which, together with Equations 4.15, 4.16 and 4.21, define the equilibrium quantities.[10]

The two ride-hailing platforms then compete by simultaneously setting prices to both sides of the market (p_1, q_1 and p_2, q_2) attempting to maximize profit ($(p_1 - q_1)x_1$ and $(p_2 - q_2)x_2$).

Numerical results

In order to obtain numerical results for the previous differentiated duopoly settings, I will use the parameter estimates from Chapter 3, which were calibrated to the morning peak period of Bogotá in 2019 (see Section 3.3). However, the bivariate distribution of reservation values and values of time $f(\beta, V)$ must be extended to a trivariate distribution $f(\beta, V_1, V_2)$.

The bivariate distribution estimated in Chapter 3 has a bivariate normal shape with parameters (point estimates): $\mu_\beta = \$9,100/hr$; $\sigma_\beta = \$5,220/hr$; $\mu_V = \$12,450$; $\sigma_V = \$5,270$ and $\rho = 0.63$.[11] I extend this bivariate normal distribution to a trivariate normal distribution by maintaining the means and standard deviations of the value of time and reservation values, as well as the coefficient of correlation between value of time and reservation values. The coefficient of correlation between the reservation values for the two platforms ($\rho_{V_1 V_2}$) determines the degree of differentiation between platforms. In one extreme, perfect positive correlation ($\rho_{V_1 V_2} = 1$) leads to undiffer-

[10] Probabilities pr_1 and pr_2 in Equations 4.27 and 4.28 have analytical expressions similar to those derived in Appendix G, but they consider, besides the difference in prices ($p_1 - p_2$), the difference in reservation values for each traveler ($V_2 - V_1$).

[11] All monetary values are in colombian pesos. The purchasing power parity (PPP) adjusted conversion rate between colombian pesos and U.S. dollars in 2019 was $1,340.5 COL/USD (OECD, 2019).

entiated platforms. On the other extreme, perfect negative correlation ($\rho_{V_1V_2} = -1$) leads to the usual Hotelling setting of product differentiation (i.e. travelers with high reservation values for one platform have low reservation values for the other). In general, stronger correlation between reservation values implies less differentiation between platforms, which leads to stronger competition. Note that the differentiation introduced by this trivariate distribution is neutral between platforms. None of the platforms has an advantage in terms of being preferred by more travelers, nor in terms of the strength of these preferences.

The trivariate distribution $f(\beta, V_1, V_2)$ that I will use in this section has then a trivariate normal shape with parameters: $\mu_\beta = \$9,100/hr$; $\sigma_\beta = \$5,220/hr$; $\mu_{V_1} = \mu_{V_2} = \$12,450$; $\sigma_{V_1} = \sigma_{V_2} = \$5,270$ and $\rho_{\beta V_1} = \rho_{\beta V_2} = 0.63$.[12] In order to choose a coefficient of correlation between reservation values, I will initially explore the effect of this coefficient on the equilibrium price charged by platforms when travelers decide between platforms based on expected wait times.[13] The numerical results show that the equilibrium reached by the two competing platforms is symmetric ($p_1 = p_2$ and $q_1 = q_2$) for any correlation.[14] Figure 4.3 graphs the equilibrium price charged by platforms at varying degrees of differentiation. As expected, less differentiation (higher correlation) leads to stronger competition and consequently to lower equilibrium prices.

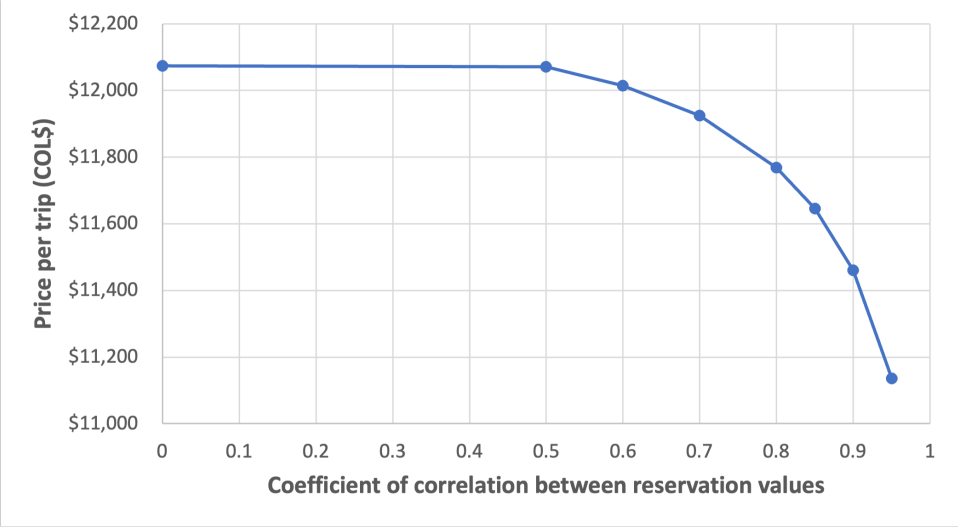
The price charged to riders by a monopolistic platform was \$11,984 (Table 4.1). This price is higher than the average price paid by ride hailers during the morning peak period in Bogotá, which was about \$11,600 (see Section 3.3.1). For the rest of

[12] As in Chapter 3, I limit the support of the distribution from \$0 to \$30,000 in reservation values and from \$100 to \$25,000/hr in values of time. The population size is $N = 145,000$.

[13] I do not have information to estimate directly the degree of differentiation between the two main ride-hailing platforms in Bogotá in 2019 (Uber 70%, Beat 25%), or more in general to estimate the trivariate distribution $f(\beta, V_1, V_2)$. The stated-preference surveys used in Chapter 3 to estimate the bivariate distribution $f(\beta, V)$ did not differentiate among the platforms available at the moment in Bogotá.

[14] I compute the equilibrium between platforms by iterative best responses starting from monopolistic pricing by one of the platforms. Only a few iterations are usually needed to achieve equilibrium, which is invariant to the initial pricing point.

Figure 4.3: Price charged by platforms to riders at varying degrees of differentiation.



Notes: In equilibrium the two platforms charge riders the same price per trip. It is assumed that riders choose between platforms based on expected wait times. Prices apply to an average-distance trip (7.66km).

the numerical results presented in this section, I will use a coefficient of correlation between reservation values $\rho_{V_1V_2} = 0.85$, which is the coefficient that lowers the price charged to riders to about \$11,600.[15]

Expected vs realized wait times

Table 4.2 compares the prices and outcomes reached by the two platforms in equilibrium (Duopoly) when travelers base their decisions on expected versus realized wait times. In both cases, the equilibrium reached by the two competing platforms is symmetric ($p_1 = p_2$ and $q_1 = q_2$). Table 4.2 also presents the prices imposed and the outcomes reached by a social planner in charge of both platforms (Welfare Max.).[16] Again, the social planner imposes symmetric prices.

The move from expected to realized wait times changes the results for the duopoly

[15]Even though this degree of differentiation leads to the right price, it does not reflect the fact that Uber controlled a higher market share than Beat (70% vs 25%). In equilibrium, both platforms achieve a 50% market share.

[16]The social planner attempts to maximize overall welfare by choosing the prices of the two platforms. The welfare function in this case is a direct generalization of the welfare functions presented in Equations 4.6 and 4.14 for a single platform.

Table 4.2: Comparison of duopoly results based on expected to realized wait times.

		Duopoly		Welfare Max.	
		Expected	Realized	Expected	Realized
Prices	Riders (p)	\$11,646	\$11,634	\$13,656	\$13,597
	Drivers (q)	\$10,043	\$10,011	\$10,132	10,059
Platf. Commission		13.76%	13.95%	25.81%	26.02%
Quantities	Trips per hour (x)	17,990	18,112	6,954	7,102
	Vehicles (d)	12,838	12,882	5,006	5,076
In-vehicle time increase		6.43%	6.45%	2.51%	2.54%
Average wait time (min)		2.87	2.73	4.42	4.33
Profit per platform (millions/hr)		\$14.42	\$14.70	\$12.25	\$12.56
Overall welfare (millions/hr)		\$-2.01	\$-1.33	\$10.04	\$10.56

Notes: All monetary values are in colombian pesos. PPP adjusted conversion rate \$1,340.5 COL/USD. Prices apply to an average-distance trip (7.66km).

and social planner scenarios in a manner similar to how it changed the results for a single platform (Table 4.1). Once travelers base their decisions on realized wait times, the prices charged to riders and paid to drivers by the platforms decrease slightly, which causes a small increase in the size of the market. The main effect of moving to realized wait times is to decrease the average wait time experienced by riders, which increases the overall welfare created by ride hailing and allows the private platforms to achieve higher profits.

These results do not support the idea that allowing travelers to check wait times before they decide to hail a ride increases competition between platforms. The main effect of this feature of modern ride-hailing platforms is simply to improve the efficiency of ride-hailing markets by decreasing the probability of long-distance matches.

In comparison to the outcomes achieved by a monopolist platform (Table 4.1), the duopoly settings increase the size of the ride-hailing market (as measured by the number of riders or drivers) by more than 60%. This significant increase exacerbates

the impact of ride hailing on traffic congestion. As a result, the availability of ride hailing actually decreases overall welfare in the duopoly setting. On the contrary, the social planner limits the growth of the market to less than 25% when moving from one to two platforms, and manages to increase the welfare created by ride-hailing services.

Optimal regulation

As in the monopolistic scenario of Chapter 3, it is not necessary for a social planner to take control of the platforms to improve the welfare created by ride hailing. Table 4.3 shows that by imposing an optimal tax on the price charged by platforms to riders, a regulator can achieve over 99% of the welfare gains available from the unregulated scenario.[17] The size of the optimal tax is $2,280.[18] This tax is 34% higher than the optimal tax in the monopolistic scenario ($1,700, Table 3.3). As a result of the tax, the price platforms collect from riders (net of tax) decreases from $11,646 to $11,450. The price paid by riders then increases by only $2,084 when the tax is introduced. This result implies a pass-trough of 0.91, which is higher than the monopolistic pass-trough from Chapter 3 (0.81).

It is important to note that the optimal tax reduces the profit of the private platforms by almost 70%. If the resulting profit is too low for a platform to cover its fixed costs (e.g. administrative costs), the tax may have the unintended consequence of changing the structure of the market from a duopoly to a monopoly, in which case the size of the tax would probably be excessive.

[17] The results of Table 4.3 assume that travelers base their decisions on expected wait times. The optimal tax and associated outcomes change only slightly when travelers consider realized wait times.

[18] This tax corresponds to the optimal charge for an average-distance trip (7.66km). The optimal tax for trips of different length should be adjusted proportionally to distance. The optimal tax should then be interpreted as $298 per kilometer.

Table 4.3: Optimal regulation of a duopoly.

		Duopoly	Social planner	Tax=$2,280
Prices	Riders (p)	$11,646	$13,656	$11,450
	Drivers (q)	$10,043	$10,132	$10,193
Platf. Commission		13.76%	25.81%	10.98%
Quantities	Trips per hour (x)	17,990	6,954	6,946
	Vehicles (d)	12,838	5,006	5,030
In-vehicle time increase		6.43%	2.51%	2.52%
Average wait time (min)		2.87	4.42	3.94
Profit per platform (millions/hr)		$14.42	$12.25	$4.37
Overall welfare (millions/hr)		$-2.01	$10.04	$9.95

Notes: All monetary values are in colombian pesos rounded to the nearest ten. PPP adjusted conversion rate $1,340.5 COL/USD. Prices apply to an average-distance trip (7.66km). The price faced by riders in the last column is COL$13,730 (=$11,450+$2,280). The results assume riders base their decisions on expected wait times.

4.3.3 Choosing the number of vehicles

In their usual business model, ride-hailing platforms set prices to both sides of the market, while riders and drivers decide to enter the market based on these prices and the number of people on the other side. The advent of autonomous vehicles, however, may change the way ride-hailing platforms operate, as they will probably choose directly the number of vehicles to have in service at any time. This change will alter the nature of competition between platforms, as they will choose quantities instead of prices in one side of the market (i.e. from Bertrand to Cournot competition). This section briefly explores the effect of this change on the equilibrium results of the differentiated duopoly settings introduced in the previous section.

In terms of the model, each of the platforms now chooses the price charged to riders (p) and the number of vehicles (d) in order to maximize profit, which equals

total revenue minus operational costs

$$\max_{p,d} \quad p \cdot x - c \cdot d \tag{4.29}$$

where x is again the number of riders and c is the operational cost of autonomous vehicles per unit of time. The same set of equations as in Section 4.3.2 determines the number of riders, as well as in-vehicle and wait times, on each platform for a given set of prices and vehicles (p_1, d_1 and p_2, d_2) chosen by the platforms.[19] Even though one of the main impacts of autonomous vehicles will be to reduce the operational cost of ride hailing, as drivers' net earnings usually represent about 75% of this cost (J. V. Hall & Krueger, 2018), I will set the operational cost of autonomous vehicles equal to the reservation wage of drivers (which includes net earnings and operational costs) in order to isolate the effect of autonomous vehicles on the nature of competition between platforms.

Table 4.4 compares the equilibrium results of duopoly competition when platforms choose prices (p, q) to when they choose price and vehicles (p, d). When platforms choose directly the number of vehicles, the equilibrium price charged to riders increases, the size of the market decreases, and platforms are able to achieve higher profits. In a few words, the strength of competition in the market decreases, as may have been expected from the basic duopoly results of Bertrand competition in prices and Cournot competition in quantities.

Even though a less competitive market usually implies lower efficiency, this is not the case. Overall welfare *increases* once platforms choose vehicles directly. As we know, this counterintuitive result arises from the presence of negative externalities. Since the optimal size of the market is smaller due to the existence of negative externalities, the market reduction caused by having platforms choose vehicles directly is

[19] Equations 4.15, which determine the number of vehicles as a function of the price per trip paid to drivers, are no longer needed.

Table 4.4: Comparison between competition in prices and competition choosing vehicles for a duopoly.

		Competition in prices (p, q)	Competition choosing vehicles (p, d)
Prices	Riders (p)	\$11,634	\$11,821
	Drivers (q)	\$10,011	-
Platf. Commission		13.95%	-
Quantities	Trips per hour (x)	18,112	16,852
	Vehicles (d)	12,882	11,990
In-vehicle time increase		6.45%	6.01%
Average wait time (min)		2.73	2.79
Profit per platform (millions/hr)		\$14.70	\$15.22
Overall welfare (millions/hr)		\$-1.33	\$-0.18

Notes: All monetary values are in colombian pesos rounded to the nearest ten. PPP adjusted conversion rate \$1,340.5 COL/USD. Prices apply to an average-distance trip (7.66km). The results assume riders base their decisions on realized wait times.

beneficial.

The overall welfare created by ride hailing when platforms choose prices directly is still far from the highest possible, so a congestion charge is still justified.[20] The optimal size of the congestion charge, as a tax on the price per trip charged to riders, is now \$2,070, which is about 10% lower than the optimal tax when platform compete in prices. As usual, this optimal charge achieves almost all of the welfare gains available from the unregulated scenario, even though the supply side of vehicles is not directly regulated.

[20]The outcomes reached by a social planner who manages the two platforms by choosing the prices charged to riders and the number of vehicles are the same as those reached when the decision variable are prices on both sides of the market. So the highest welfare possible is equal to that of Table 4.3 (\$10.04 millions/hr).

4.4 Conclusions

The results obtained in this chapter have a common theme. Because the distortion caused by negative traffic congestion externalities is larger than the distortion caused by market power for the ride-hailing market under analysis (morning peak period of Bogotá), increases in market size caused by a more competitive market reduce overall welfare. The strength of competition in ride-hailing markets can change due to the addition of differentiated platforms (more competition), or due to platforms choosing directly the number of vehicles to have available for service (less competition). The fact that modern platforms allow riders to check the location of the closest idle vehicle before deciding to hail a ride does not change significantly the strength of competition in the market. The main impact of this technological feature is to reduce average wait times by improving the efficiency of rider-driver matching.

However, the distortion created by traffic congestion externalities is not necessarily larger than market power in all ride-hailing markets. Recall that Bogotá is usually ranked as one of the most congested cities in the world. Market power may be the dominant distortion in other less congested cities, especially during off-peak periods, in which case a more competitive ride-hailing market would lead to welfare gains.

Appendix A

Derivation of formulas for the optimal and revenue-maximizing access fees

This appendix derives the formulas for the optimal and revenue-maximizing access fees presented in Equations 2.4 and 2.7 of Sections 2.2.3 and 2.2.4 respectively.

A.1 Optimal fee

Expression 2.3 presents the objective function of the social planner's welfare maximization problem as a function of the access fee τ. Differentiate this objective function with respect to τ to obtain the following first-order condition (the optimal number of users and access fee were denoted x^* and τ^* in the body of the article, but I will denote them just as x and τ in this appendix to facilitate notation)

$$\begin{aligned} &- g'(x)\frac{\mathrm{d}x}{\mathrm{d}\tau}\int_{\tau+\beta g(x)}^{\infty}\int_{0}^{\infty}\beta f(\beta,V)\,\mathrm{d}\beta\,\mathrm{d}V \\ &\quad - \tau\left[\int_{0}^{\infty} f(\beta,\tau+\beta g(x))\,\mathrm{d}\beta + g'(x)\frac{\mathrm{d}x}{\mathrm{d}\tau}\int_{0}^{\infty}\beta f(\beta,\tau+\beta g(x))\,\mathrm{d}\beta\right] = 0 \end{aligned} \quad \text{(A.1)}$$

The derivative of the access fee with respect to the number of users ($\mathrm{d}x/\,\mathrm{d}\tau$) can be obtained from Equation 2.2 using the implicit function theorem or total differentiation. Either way, one gets

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = -\left[\frac{N\int_{0}^{\infty} f(\beta,\tau+\beta g(x))\,\mathrm{d}\beta}{1+Ng'(x)\int_{0}^{\infty}\beta f(\beta,\tau+\beta g(x))\,\mathrm{d}\beta}\right] \quad \text{(A.2)}$$

Substitute Equation A.2 into Equation A.1, simplify and rearrange terms to obtain

$$\tau = Ng'(x) \int_{\tau+\beta g(x)}^{\infty} \int_{0}^{\infty} \beta f(\beta, V)\, \mathrm{d}\beta\, \mathrm{d}V \tag{A.3}$$

Equation 2.5 expresses mathematically the average sensitivity to congestion of all users ($\bar{\beta}$). The denominator of this expression represents the fraction of the population that are users, which multiplied by N gives the number of users x. Expression A.3 then simplifies to

$$\tau = \bar{\beta} x g'(x), \tag{A.4}$$

which is the expression presented in the body of the article.

A.2 Revenue-maximizing fee

Expression 2.6 presents the objective function of the private firm as a function of the access fee. Differentiate this objective function with respect to τ to obtain the following first-order condition (the number of users and access fee for the private firm were denoted x^p and τ^p in the body of the article, but I will denote them just as x and τ in this appendix to facilitate notation)

$$\tau \frac{\mathrm{d}x}{\mathrm{d}\tau} + x = 0 \tag{A.5}$$

Substitute Equation A.2 into Equation A.5 and rearrange terms to obtain

$$\tau = x \left[\frac{1 + Ng'(x) \int_{0}^{\infty} \beta f(\beta, \tau + \beta g(x))\, \mathrm{d}\beta}{N \int_{0}^{\infty} f(\beta, \tau + \beta g(x))\, \mathrm{d}\beta} \right] \tag{A.6}$$

Equation 2.8 expresses mathematically the average sensitivity to congestion of

marginal users. Using this expression, the previous equation simplifies to

$$\tau = \bar{\beta}_m x g'(x) + \frac{x}{N \int_0^\infty f(\beta, \tau + \beta g(x))\, \mathrm{d}\beta} \tag{A.7}$$

Note that if congestion is fixed at a given level $\bar{g}$, the derivative of the number of users with respect to the access fee ($\mathrm{d}x/\,\mathrm{d}\tau$, Equation A.2) reduces to

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = -N \int_0^\infty f(\beta, \tau + \beta \bar{g})\, \mathrm{d}\beta, \tag{A.8}$$

and the elasticity of demand can be expressed as

$$\bar{\varepsilon} = -\frac{\tau}{x} N \int_0^\infty f(\beta, \tau + \beta \bar{g})\, \mathrm{d}\beta \tag{A.9}$$

Substitute this expression into Equation A.7 to obtain

$$\tau = \bar{\beta}_m x g'(x) + \frac{\tau}{\bar{\varepsilon}}, \tag{A.10}$$

which is the expression presented in the body of the article.

Appendix B

The scale-income model

The main purpose of this appendix is to show that if the distribution of sensitivities to congestion across potential users follows a Pareto distribution, the scale-income model implies that the revenue-maximizing access fee equals the optimal fee, so private management of the congestible resource leads to an efficient level of use. As discussed in Section 2.2.6, the scale-income model can be introduced by assuming that all potential users have as their outside option an alternative with a constant level of congestion. For concreteness, I will use in this appendix the example of travelers choosing between a congestible and an uncongestible road, which is actually the example originally used by Pigou (1920).

Travel time in the congestible road is an increasing function of the number of travelers $t(x)$, while travel time in the uncongestible road is fixed $\bar{t}$.[1] Travelers choose the road that minimizes their total cost, which includes the cost of travel time plus any potential toll τ on the congestible road. For traveler i with value of time β_i, total cost on the congestible road is $\tau + \beta_i \cdot t(x)$, while total cost on the uncongestible road is $\beta_i \cdot \bar{t}$.

The univariate density function $f(\beta)$, with associated cumulative function $F(\beta)$, represents the distribution of values of time in the population of travelers. All travelers have a positive value of time. The total number of travelers is N. Given a positive

[1] I assume $t(0) < \bar{t}$. Otherwise, the example lacks any interest.

toll ($\tau > 0$), the Nash Equilibrium (NE) allocation of travelers between the two roads involves travelers with the highest values of time choosing the congestible road.[2] A specific value of time β_m then characterizes the NE allocation of travelers. Travelers with values of time above β_m choose the congestible road, while travelers with lower values of time choose the uncongestible one. Travelers with value of time β_m are indifferent between the two roads. The following condition determines the number of travelers on the congestible road (x) as a function of the toll (τ)

$$\beta_m \cdot \bar{t} = \tau + \beta_m \cdot t(x) \tag{B.1}$$

where β_m and x are related by $x = (1 - F(\beta_m))N$.

Equation B.1 clearly implies that $\bar{t} > t(x)$, so the congestible road remains faster than the uncongestible one. Note that all travelers with value of time below β_m strictly prefer the uncogestible road, while travelers with value of time above β_m strictly prefer the congestible road. None of the travelers has an incentive to switch roads, satisfying the condition for a NE.

In this example, the objective of a social planner is to choose the value of the toll to minimize the travel time cost of all travelers, while the objective of a private firm is to maximize toll revenue. Following an approach similar to that employed in Appendix A, it can be shown that the conditions for the optimal and revenue-maximizing tolls are

$$\tau^* = \bar{\beta}^* x^* t'(x^*) \tag{B.2}$$

$$\tau^p = \beta_m^p x^p t'(x^p) + (\bar{t} - t(x^p)) \frac{1 - F(\beta_m^p)}{f(\beta_m^p)} \tag{B.3}$$

where $\bar{\beta}^*$ denotes the average value of time of travelers on the congestible road under

[2] To avoid the corner solutions where all travelers choose one of the roads, I assume $\beta_{min}(\bar{t} - t(N)) < \tau < \beta_{max}(\bar{t} - t(0))$, where β_{min} and β_{max} are the lowest and highest values of time among travelers.

the optimal allocation, and β_m^p represents the value of time of travelers indifferent between the two roads under the revenue-maximizing allocation (marginal travelers). These two conditions have the same form and interpretation as those in Appendix A.

We can use Equation B.1 to substitute τ out of the previous conditions and express them as

$$\beta_m^*(\bar{t} - t(x^*)) = \bar{\beta}^* x^* t'(x^*) \tag{B.4}$$

$$(\bar{t} - t(x^p))\frac{1 - F(\beta_m^p)}{f(\beta_m^p)} - \beta_m^p(\bar{t} - t(x^p) - x^p t'(x^p)) = 0 \tag{B.5}$$

The second-order condition of the revenue maximization problem implies that the left-hand side of Equation B.5 is increasing in x. Consequently, if the left-hand side of Equation B.5 is positive when evaluated at x^*, we can conclude that $x^* > x^p$, which implies that the optimal toll is lower than the revenue-maximizing toll ($\tau^* < \tau^p$). Similarly, if Equation B.5 holds at x^*, the revenue-maximizing toll equals the optimal one.[3] Evaluating Equation B.5 at x^* and using Equation B.4, the resulting condition for the revenue-maximizing toll to be optimal can be expressed as

$$\frac{1 - F(\beta_m^*)}{f(\beta_m^*)} - \frac{\beta_m^*}{\bar{\beta}^*}(\bar{\beta}^* - \beta_m^*) = 0 \tag{B.6}$$

The term $(\bar{\beta}^* - \beta_m^*)$ corresponds to the mean residual life of the distribution of values of time evaluated at β_m^*, while the term $\frac{1-F(\beta_m^*)}{f(\beta_m^*)}$ corresponds to the inverse of the hazard or failure rate. Denote the mean residual life function by $m(\beta)$ and the hazard rate function by $h(\beta)$. The mean residual life and hazard rate of any distribution are related by $m(\beta)h(\beta) = 1 + m'(\beta)$, where $m'(\beta)$ is the derivative of the mean residual life function. After some slight manipulation, Equation B.6 can

[3] I assume the revenue maximization problem does not have multiple local maxima (quasi-concavity).

then be transformed to

$$1 - \frac{1 + m'(\beta_m^*)}{1 + m(\beta_m^*)/\beta_m^*} = 0 \tag{B.7}$$

The Pareto distribution has the special property that $m'(\beta) = m(\beta)/\beta$. It is then clear from Equation B.7 that if values of time (sensitivities to congestion in general) follow a Pareto distribution, the revenue-maximizing toll would be socially optimal and private ownership would lead to an efficient level of use of the congestible road. Equation B.7 allows us to say a bit more. Any distribution with a non-decreasing hazard rate satisfies $m'(\beta) < 0$ (e.g. normal or exponential distributions). In such cases, the left-hand side of Equation B.7 is positive and a private firm in charge of the congestible road imposes a toll above the optimal level ($\tau^* < \tau^p$). Only distributions with heavy tails, such as Pareto, can induce the private firm to impose a toll below the optimal level.

Appendix C

Expected wait time

This appendix provides a simple proof of the expected wait time formula presented in Equation 3.5. This formula was first derived by Arnott (1996). The proof I provide here uses an approach different from that used by Arnott.

The expected wait time formula states that given a density of idle vehicles D, the expected wait time for a rider is $\frac{1}{2v\sqrt{D}}$, where v is the speed of vehicles when en route to pick up a passenger. This formula implies that the expected distance from the rider to the closest idle vehicle is $\frac{1}{2\sqrt{D}}$. The substance of the proof is to demonstrate this last statement.

Picture a rider at the center of a circular area of size A. If I idle vehicles were located randomly inside this area, to reach density $D = I/A$, the probability that all vehicles are at a distance greater than y from the rider is

$$pr(r \geq y) = \left(1 - \frac{\pi y^2}{A}\right)^I \tag{C.1}$$

were r is the distance from the rider to the closest idle vehicle.

If the size of the area is relatively large in comparison to the expected distance to the closest vehicle, we can approximate the previous expression assuming A goes to infinity while the density of idle vehicles remains constant at D. First, replace

$I = AD$ in the previous equation to obtain

$$pr(r \geq y) = \left(1 - \frac{\pi y^2}{A}\right)^{AD} \tag{C.2}$$

The limit of the previous expression as A goes to infinity is

$$pr(r \geq y) = e^{-\pi y^2 D} \tag{C.3}$$

The negative of the derivative of the previous expression with respect to y evaluated at r gives the density function of the distance to the closest idle vehicle. Denoting this density function by $f(r)$, we have

$$f(r) = 2\pi r D e^{-\pi r^2 D} \tag{C.4}$$

The expected distance to the closest idle vehicle can then be computed as

$$E(r) = \int_0^\infty r f(r)\,\mathrm{d}r = \int_0^\infty 2\pi r^2 D e^{-\pi r^2 D}\,\mathrm{d}r \tag{C.5}$$

The previous integral can be simplified by the change of variable $a = \pi r^2$, which leads to

$$E(r) = \int_0^\infty \sqrt{\frac{a}{\pi}} D e^{-aD}\,\mathrm{d}a = \frac{D}{\sqrt{\pi}} \int_0^\infty \sqrt{a} e^{-aD}\,\mathrm{d}a \tag{C.6}$$

The value of the last integral is $\frac{\pi}{2D^{3/2}}$,[1] which leads to the desired result

$$E(r) = \frac{1}{2\sqrt{D}} \tag{C.7}$$

[1] I used the integral calculator available at www.integral-calculator.com.

Appendix D

Equilibrium number of riders and drivers

This appendix explores the possible solutions to Equations 3.2 and 3.4, which determine the equilibrium number of riders and drivers (x, d) for any given set of prices charged to riders and paid to drivers per trip (p, q). For convenience, I rewrite these two equations here

$$N \int_{\beta=0}^{\infty} \int_{V=p+\beta(t(d)+w(x,d))}^{\infty} f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta = x \tag{D.1}$$

$$d = \frac{q}{c} \cdot x \tag{D.2}$$

where $w(x, d)$ (expected wait time as a function of the number of riders an drivers) in Equation D.1 is given by (see Equation 3.8 in Section 3.2.3)

$$w(x, d) = \frac{1}{2v\sqrt{\frac{d - s \cdot x}{A}}} \tag{D.3}$$

an $t(d)$ (in-vehicle travel time as a function of the number of vehicles) is a linear function that describes the effect of ride-hailing vehicles on traffic congestion.

To analyze the potential solutions to Equations D.1 and D.2, it will prove convenient to first do a slight algebraic manipulation. Use Equation D.2 to substitute x out of Equation D.3 and express expected wait time as a function of only the number

of drivers

$$w(d) = \frac{1}{2v\sqrt{\frac{(1-c\cdot s/q)d}{A}}} \tag{D.4}$$

We can now insert this expression for expected wait time into the left-hand side of Equation D.1 to express the number of riders as a function of the number of drivers and the price charged to riders per trip

$$x(d;p) = N\int_0^\infty \int_{p+\beta(t(d)+w(d))}^\infty f(\beta, V)\,\mathrm{d}V\,\mathrm{d}\beta \tag{D.5}$$

Recall also that Equation D.2 determines the number of drivers as a function of the number of riders and the price paid to drivers per trip

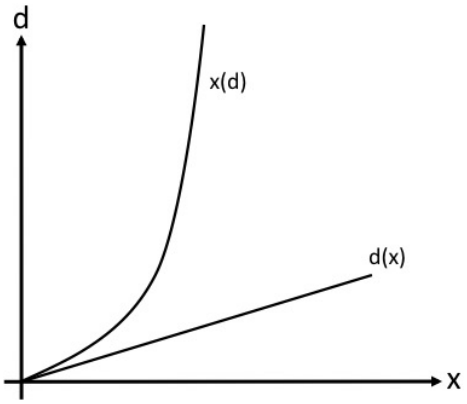
$$d(x;q) = \frac{q}{c}\cdot x \tag{D.6}$$

Equations D.5 and D.6 allow us to examine the potential equilibrium points in a two-dimensional graph with x in the horizontal axis and d in the vertical one (Figure D.1). Equation D.6 represents a straight line that goes through the origin. The slope of this line increases with q (the price per trip paid to drivers). Equation D.5 describes a function that is increasing at low values of d, but may become decreasing as the effect of vehicles on traffic congestion outweighs the reduction in wait time. Since Equation D.4 is convex, Equation D.5 tends to be concave, but its second derivative depends also on the shape of the bivariate distribution $f(\beta, V)$. Increases in the price charged to riders (p) shift the function described by Equation D.5 downward (to the left in Figure D.1).

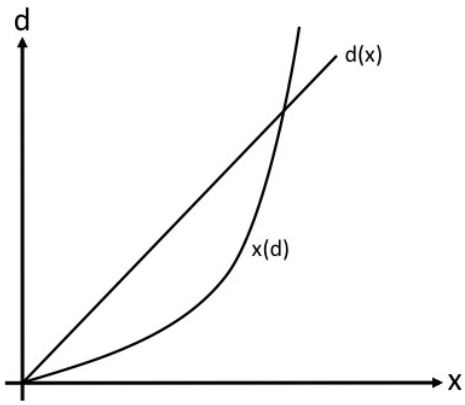
If Equation D.5 is always concave, there may be at most two equilibria. The trivial equilibrium with $(x, d) = (0, 0)$ is always present and may be the only one (Figure D.1a). However, as the price charged to riders decreases, or the price paid to drivers increases, there may exist a second equilibrium with positive numbers of

riders and drivers (Figure D.1b). If Equation D.5 is not always concave, there may exist several equilibria with positive numbers of riders and drivers (Figure D.1c).

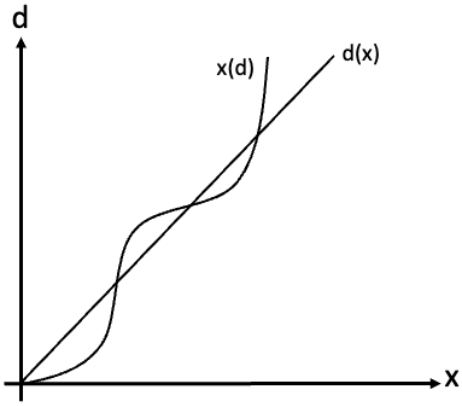
Figure D.1: Equilibrium number of riders (x) and drivers (d).



(a) One equilibrium.



(b) Two equilibria.



(c) Multiple equilibria.

Note: Panel (a) illustrates a situation where the only equilibrium has no drivers and no riders. In panel (b), there is one additional equilibrium with positive numbers of riders and drivers. Panel (c) contains several equilibria with positive riders and drivers.

Appendix E

Profit- vs Welfare-maximizing prices

This appendix derives the expressions for the welfare- and profit-maximizing price gaps analyzed in section 3.2.6 (Equations 3.12 and 3.14). It also presents the additional first-order condition that is needed in each problem to determine both prices. The methodology used in this appendix is very similar, but slightly more complex, than that used in Appendix A.

E.1 Welfare maximization

The welfare maximization problem can be stated mathematically as (section 3.2.6):

$$\max_{p,q} \quad N \int_0^\infty \int_{p+\beta(t(d)+w(x,d))}^\infty [V - \beta(t(d) + w(x,d))] f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta - c \cdot d - MEC \cdot d \tag{E.1}$$

where x and d are implicit functions of p and q through the following two equations (sections 4.1 and 4.2):

$$N \int_0^\infty \int_{p+\beta(t(d)+w(x,d))}^\infty f(\beta, V)\, \mathrm{d}V\, \mathrm{d}\beta = x \tag{E.2}$$

$$d = \frac{q}{c} x \tag{E.3}$$

We can take d out of the problem using Equation E.3. The maximization problem

then turns to:

$$\max_{p,q} \quad N \int_0^{\infty} \int_{p+\beta\left(t\left(\frac{q}{c}x\right)+w\left(x,\frac{q}{c}x\right)\right)}^{\infty} \left[V - \beta\left(t\left(\frac{q}{c}x\right) + w\left(x,\frac{q}{c}x\right)\right)\right] f(\beta,V)\,\mathrm{d}V\,\mathrm{d}\beta \\ - q\cdot x - MEC\cdot\frac{q}{c}x \quad \text{(E.4)}$$

while x becomes an implicit function of p and q through the following equation:

$$N \int_0^{\infty} \int_{p+\beta\left(t\left(\frac{q}{c}x\right)+w\left(x,\frac{q}{c}x\right)\right)}^{\infty} f(\beta,V)\,\mathrm{d}V\,\mathrm{d}\beta = x \quad \text{(E.5)}$$

The first order condition of problem E.4 with respect to p is:

$$-\frac{\mathrm{d}x}{\mathrm{d}p}\left(t_d\frac{q}{c} + w_x + x_d\frac{q}{c}\right) N \int_0^{\infty} \int_{p+\beta(t+w)}^{\infty} \beta f(\beta,V)\,\mathrm{d}V\,\mathrm{d}\beta \\ - N\int_0^{\infty} p\left[1+\beta\frac{\mathrm{d}x}{\mathrm{d}p}\left(t_d\frac{q}{c} + w_x + x_d\frac{q}{c}\right)\right] f(\beta, p+\beta(t+w))\,\mathrm{d}\beta \\ - q\frac{\mathrm{d}x}{\mathrm{d}p} - MEC\frac{q}{c}\frac{\mathrm{d}x}{\mathrm{d}p} = 0 \quad \text{(E.6)}$$

where t_d, w_x and w_d denote the partial derivatives of the in-vehicle and wait time functions with respect to the number of riders and vehicles. From equation E.5 and the implicit function theorem, the derivative of the number of riders with respect to price is:

$$\frac{\mathrm{d}x}{\mathrm{d}p} = \frac{-N\int_0^{\infty} f(\beta, p+\beta(t+w))\,\mathrm{d}\beta}{1 + \left(t_d\frac{q}{c} + w_x + x_d\frac{q}{c}\right) N\int_0^{\infty} \beta f(\beta, p+\beta(t+w))\,\mathrm{d}\beta} \quad \text{(E.7)}$$

Plugging Expression E.7 into Equation E.6 and noting that the average value of time of riders can be expressed as:

$$\bar{\beta} = \frac{\int_0^{\infty}\int_{p+\beta(t+w)}^{\infty} \beta f(\beta,V)\,\mathrm{d}V\,\mathrm{d}\beta}{\int_0^{\infty}\int_{p+\beta(t+w)}^{\infty} f(\beta,V)\,\mathrm{d}V\,\mathrm{d}\beta} \quad \text{(E.8)}$$

the first-order condition can be written as:

$$p = \bar{\beta} x \left(t_d \frac{q}{c} + w_x + x_d \frac{q}{c} \right) + q + \frac{q}{c} MEC \tag{E.9}$$

Expressing the term in parentheses as $\mathrm{d}(t+w)/\mathrm{d}x$, the (not partial) derivative of travel time with respect to the number of drivers, noting that $q/c = d/x$ from Equation E.3 and moving q to the left-hand side, one obtains Equation 3.12, which is the one analyzed in section 3.2.6.

The first-order condition for q, after a similar but slightly more cumbersome process, can be expressed as:

$$c + \bar{\beta} \cdot x \cdot t_d + MEC = -\bar{\beta} \cdot x \cdot w_d \tag{E.10}$$

We can interpret this condition as follows. The left-hand side reveals the cost of having one more ride-hailing vehicle, which includes the direct cost of vehicle expenses and driver labor (c), the external congestion cost on ride hailers ($\bar{\beta} x t_d$) and the external congestion cost on other road users (MEC). The right-hand side measures the benefit of that additional vehicle, which is the reduction in wait time for ride hailers ($-\bar{\beta} x w_d$). At an optimal solution, these marginal costs and benefit must be equal.

E.2 Profit maximization

The profit maximization problem is relatively simpler:

$$\max_{p,q} \quad (p - q) \cdot x \tag{E.11}$$

while x is again an implicit function of p and q through Equation E.5.

The first order condition of problem E.11 with respect to p is:

$$x + (p - q) \cdot \frac{\mathrm{d}x}{\mathrm{d}p} = 0 \tag{E.12}$$

Plugging Expression E.7 into Equation E.12 and noting that the average value of time of *marginal* riders can be expressed as:

$$\bar{\beta}_m = \frac{\int_0^\infty \beta f(\beta, p + \beta(t+w))\,\mathrm{d}\beta}{\int_0^\infty f(\beta, p + \beta(t+w))\,\mathrm{d}\beta} \tag{E.13}$$

the first-order condition can be written as:

$$p = \bar{\beta}_m x \left(t_d \frac{q}{c} + w_x + x_d \frac{q}{c} \right) + \frac{x}{N \int_0^\infty f(\beta, p + \beta(t+w))\,\mathrm{d}\beta} + q \tag{E.14}$$

Expressing the term in parentheses as $\mathrm{d}(t+w)/\mathrm{d}x$ (as in the welfare-maximization problem), we move to:

$$p = \bar{\beta}_m x \frac{\mathrm{d}(t+w)}{\mathrm{d}x} + \frac{x}{N \int_0^\infty f(\beta, p + \beta(t+w))\,\mathrm{d}\beta} + q \tag{E.15}$$

From Equation E.5 (or more directly from Expression E.7), the derivative of the number of riders with respect to price when travel time $(t+w)$ is taken as fixed is:

$$\frac{\mathrm{d}x}{\mathrm{d}p}\Big|_{t+w} = -N \int_0^\infty f(\beta, p + \beta(t+w))\,\mathrm{d}\beta \tag{E.16}$$

The absolute value of the elasticity of the number of riders with respect to price, again holding travel time fixed, can then be expressed as:

$$\varepsilon = \frac{p}{x} \cdot N \int_0^\infty f(\beta, p + \beta(t+w))\,\mathrm{d}\beta \tag{E.17}$$

Using this last expression to simplify the second term on the right-hand side of Equa-

tion E.15, and moving q to the left-hand side, one obtains Equation 3.14, which is the one analyzed in section 3.2.6.

After a similar process, the first-order condition for q can be expressed as:

$$c + \bar{\beta}_m \cdot x \cdot t_d = -\bar{\beta}_m \cdot x \cdot w_d \tag{E.18}$$

Comparing this condition to the equivalent one for welfare maximization (Equation E.10), we note that the profit-maximizing platform does not take into account the external congestion cost on other road users. Additionally, it values the external congestion cost on ride hailers using the average value of time of *marginal* riders (Spence distortion). As mentioned on section 4.2, there is no markdown distortion on this side of the market (the side of drivers) due to the uniform-reservation-wage assumption.

Appendix F

Conditional expected wait time

This appendix derives the formula for the expected wait time faced by a traveler who chooses to hail a ride only if the closest idle vehicle is at a distance shorter than y (Equation 4.13). By differentiation of Expression 4.7, the probability density function of the distance to the closest idle vehicle is

$$f(r) = 2\pi r D e^{-\pi r^2 D} \tag{F.1}$$

where D is the density of idle vehicles in the service area.

The expected distance to the closest idle vehicle, conditional on it being less than y, can be computed as

$$E[r|r \leq y] = \frac{\int_0^y r f(r)\,\mathrm{d}r}{\int_0^y f(r)\,\mathrm{d}r} \tag{F.2}$$

The integral in the denominator corresponds to the probability that the closest vehicle is at a distance shorter than y, which equals $1 - e^{-\pi y^2 D}$ (Expression 4.7). The integral in the numerator is slightly more cumbersome[1]

$$\int_0^y r f(r)\,\mathrm{d}r = \int_0^y 2\pi r^2 D e^{-\pi r^2 D} = \frac{\operatorname{erf}\left(\sqrt{\pi D} y\right)}{2\sqrt{D}} - y e^{-\pi y^2 D} \tag{F.3}$$

[1] I used the calculator available at www.integral-calculator.com.

The conditional expected distance to the closest idle vehicle is then

$$E[r|r \leq y] = \frac{\frac{\text{erf}\left(\sqrt{\pi D}y\right)}{2\sqrt{D}} - ye^{-\pi y^2 D}}{1 - e^{-\pi y^2 D}} \tag{F.4}$$

The conditional expected wait time then results from dividing the previous expression by v (the speed of vehicles en route to pick up a passenger), which leads to the formula presented in Equation 4.13.

Appendix G

Probabilities of choosing between ride-hailing platforms based on realized wait times

The objective of this appendix is to derive analytical expressions for the probabilities that a traveler with value of time β and reservation value V chooses either one of two platforms or her outside option. Platform 2 charges a higher price $p_2 > p_1$ and the densities of idle vehicles offered by each platform are D_1 and D_2. As shown in Appendix C, a density of idle vehicles D determines a probability distribution for the distance r between a traveler and the closest idle vehicle given by

$$f(r) = 2\pi r D e^{-\pi r^2 D} \tag{G.1}$$

with support $[0, \infty)$. The associated inverse cumulative density function is

$$pr(r \geq y) = e^{-\pi y^2 D} \tag{G.2}$$

Vehicles en route to pick up a passenger travel at speed v, so the wait time to the closest idle vehicle is $w = r/v$.

The traveler checks the realized wait time of each platform before making a decision. She chooses the alternative that gives her the highest value, where the value of

each option is

$$\text{Platform } 1 : V - \beta \cdot [t + w_1] - p_1$$

$$\text{Platform } 2 : V - \beta \cdot [t + w_2] - p_2$$

$$\text{Outside option} : 0$$

It is assumed that both platforms have a positive probability of being chosen, which implies that $V - \beta t - p_2 > 0$.

Denote by y_1 and y_2 the maximum distance to the closest idle vehicle that the traveler is willing to accept in order to prefer either platform to her outside option (clearly $y_1 > y_2$). These distances can be computed as

$$y_1 = \frac{(V - \beta t - p_1)v}{\beta}; \qquad y_2 = \frac{(V - \beta t - p_2)v}{\beta} \tag{G.3}$$

Additionally, denote by z the minimum distance differential to the closest idle vehicle in favor of Platform 2 so that the traveler prefers Platform 2 over Platform 1. This distance can be computed as

$$z = \frac{(p_2 - p_1)v}{\beta} \tag{G.4}$$

Let's start by computing the probability of choosing Platform 2 (the platform with the highest price). For Platform 2 to be the preferred option, the distance to its closest idle vehicle r_2 must lower than y_2, and the distance to the closest idle vehicle of Platform 1 r_1 must be greater than $r_2 + z$. The probability of choosing Platform 2 can then be expressed as

$$pr_2(\beta, V, p_1, p_2, D_1, D_2) = \int_0^{y_2} pr(r_1 \geq r_2 + z) f(r_2)\, \mathrm{d}r_2 \tag{G.5}$$

Using Expressions G.1 and G.2

$$pr_2(\beta, V, p_1, p_2, D_1, D_2) = \int_0^{y_2} 2\pi r_2 D_2 \exp\left(-\pi r_2^2 D_2\right) \exp\left(-\pi (r_2 + z)^2 D_1\right) \, \mathrm{d}r_2 \tag{G.6}$$

After some work, one can obtain an analytical (but not simple) expression for the previous integral[1]

$$pr_2(\beta, V, p_1, p_2, D_1, D_2) = \frac{D_2}{D_1 + D_2} \left(\exp\left(-\pi z^2 D_1\right) - \exp\left(-\pi (y_1^2 D_1 + y_2^2 D_2)\right) \right.$$
$$\left. - \frac{\pi D_1 z}{\sqrt{D_1 + D_2}} \exp\left(-\frac{\pi D_1 D_2 z^2}{D_1 + D_2}\right) \left[\operatorname{erf}\left(\frac{\sqrt{\pi}(D_1 y_1 + D_2 y_2)}{\sqrt{D_1 + D_2}}\right) - \operatorname{erf}\left(\frac{\sqrt{\pi} D_1 z}{\sqrt{D_1 + D_2}}\right) \right] \right) \tag{G.7}$$

where $\operatorname{erf}(\cdot)$ is the error function.

We can compute the probability of choosing Platform 1 in a similar manner, with the slight difference that if $r_1 < z$ Platform 1 is always preferred to Platform 2. The probability of choosing Platform 1 can be expressed as

$$pr_1(\beta, V, p_1, p_2, D_1, D_2) = \int_0^z f(r_1) \, \mathrm{d}r_1 + \int_z^{y_1} pr(r_2 \geq r_1 - z) f(r_1) \, \mathrm{d}r_1 \tag{G.8}$$

$$pr_1(\beta, V, p_1, p_2, D_1, D_2) = 1 - e^{-\pi z^2 D_1} + \int_z^{y_1} 2\pi r_1 D_1 \exp\left(-\pi r_1^2 D_1\right) \exp\left(-\pi (r_1 - z)^2 D_2\right) \, \mathrm{d}r_1 \tag{G.9}$$

[1] I used the integral calculator available at www.integral-calculator.com.

$$pr_1(\beta, V, p_1, p_2, D_1, D_2) = 1 - e^{-\pi z^2 D_1} + \frac{D_1}{D_1 + D_2}\left(\exp\left(-\pi z^2 D_1\right) - \exp\left(-\pi(y_1^2 D_1 + y_2^2 D_2)\right) + \frac{\pi D_2 z}{\sqrt{D_1 + D_2}} \exp\left(-\frac{\pi D_1 D_2 z^2}{D_1 + D_2}\right)\left[\text{erf}\left(\frac{\sqrt{\pi}(D_1 y_1 + D_2 y_2)}{\sqrt{D_1 + D_2}}\right) - \text{erf}\left(\frac{\sqrt{\pi} D_1 z}{\sqrt{D_1 + D_2}}\right)\right]\right) \quad \text{(G.10)}$$

The probability of choosing the outside option is much simpler. The traveler chooses the outside option if $r_1 \geq y_1$ and $r_2 \geq y_2$, so

$$pr_0(\beta, V, p_1, p_2, D_1, D_2) = pr(r_1 \geq y_1) \cdot pr(r_2 \geq y_2) = \exp\left(-\pi(y_1^2 D_1 + y_2^2 D_2)\right) \quad \text{(G.11)}$$

Note that the previous analytical expressions satisfy a few simple checks. First, it is straightforward to verify that $pr_0 + pr_1 + pr_2 = 1$. Additionally, $pr_1 = 0$ when $D_1 = 0$ and $pr_2 = 0$ when $D_2 = 0$. Finally, as $D_2 \to \infty$, $pr_2 \to e^{-\pi z^2 D_1}$ (i.e. if Platform 2 has many idle vehicles, the only chance for Platform 1 is to have an idle vehicle closer than z to the traveler); and as $D_1 \to \infty$, $pr_1 \to 1$ (i.e. if Platform 1 has many idle vehicles, Platform 2 has no chance).

Appendix H

Duopoly equilibrium without differentiation

The objective of this appendix is to show that in the duopoly settings without differentiation introduced in Section 4.3.1, if one of the platforms sets prices (p, q) and the other $(p, q + \epsilon)$, where ϵ is a positive but potentially very small amount, only the platform with the highest price paid to drivers per trip $(q + \epsilon)$ can obtain a positive number of riders and drivers in equilibrium (i.e. this platform wins the entire market). I will use the duopoly setting in which riders base their decisions on realized wait times. The line of argument for the setting based on expected wait times is similar.

Equations 4.15, 4.16, 4.21, 4.22 and 4.23 determine the equilibrium numbers of riders (x_1, x_2), drivers (d_1, d_2) and densities of idle vehicles (D_1, D_2) for each platform, as well as the in-vehicle travel time (t), for any given pairs of prices (p_1, q_1) and (p_2, q_2) set by the platforms (it is assumed that $p_2 \geq p_1$). The equilibrium analysis can be greatly simplified by using Equations 4.15 to substitute d_1 and d_2 out of Equations 4.21, and then using the resulting equations to substitute x_1 and x_2 out of Equations 4.22 and 4.23. I will further assume that in-vehicle travel time is constant.[1] The resulting two equations, which determine the equilibrium densities of idle vehicles

[1] Changes in in-vehicle travel time due to traffic congestion affect only the total size of the ride-hailing market, but not the distribution between platforms because in-vehicle travel time is always equal for the two platforms.

(D_1, D_2), are

$$\frac{A}{\frac{q_1}{c} - s} D_1 = N \int_0^{\infty} \int_{p_1+\beta t}^{p_2+\beta t} pr(\beta, V, p_1, D_1) \cdot f(\beta, V) \, \mathrm{d}V \, \mathrm{d}\beta + N \int_0^{\infty} \int_{p_2+\beta t}^{\infty} pr_1(\beta, V, p_1, p_2, D_1, D_2) \cdot f(\beta, V) \, \mathrm{d}V \, \mathrm{d}\beta \quad \text{(H.1)}$$

$$\frac{A}{\frac{q_2}{c} - s} D_2 = N \int_0^{\infty} \int_{p_2+\beta t}^{\infty} pr_2(\beta, V, p_1, p_2, D_1, D_2) \cdot f(\beta, V) \, \mathrm{d}V \, \mathrm{d}\beta \quad \text{(H.2)}$$

Equation H.1 implicitly determines D_1 as a function of D_2. Conversely, Equation H.2 implicitly determines D_2 as a function of D_1. The solution to these two equations determines the equilibrium densities of idle vehicles for the two platforms given their prices. If the equilibrium density of idle vehicles for a platform is zero, it implies the platform obtains no riders and no drivers. If the density is positive, both the number of riders and drivers are positive.

Assuming that platforms charge the same price per trip p to riders ($p_1 = p_2 = p$), and using Expressions G.9 and G.6 for the probabilities pr_1 and pr_2, the previous equations reduce to

$$\frac{A}{\frac{q_1}{c} - s} D_1 = N \int_0^{\infty} \int_{p+\beta t}^{\infty} \int_0^{y} 2\pi r D_1 \exp\left(-\pi r^2 (D_1 + D_2)\right) f(\beta, V) \, \mathrm{d}r \, \mathrm{d}V \, \mathrm{d}\beta \quad \text{(H.3)}$$

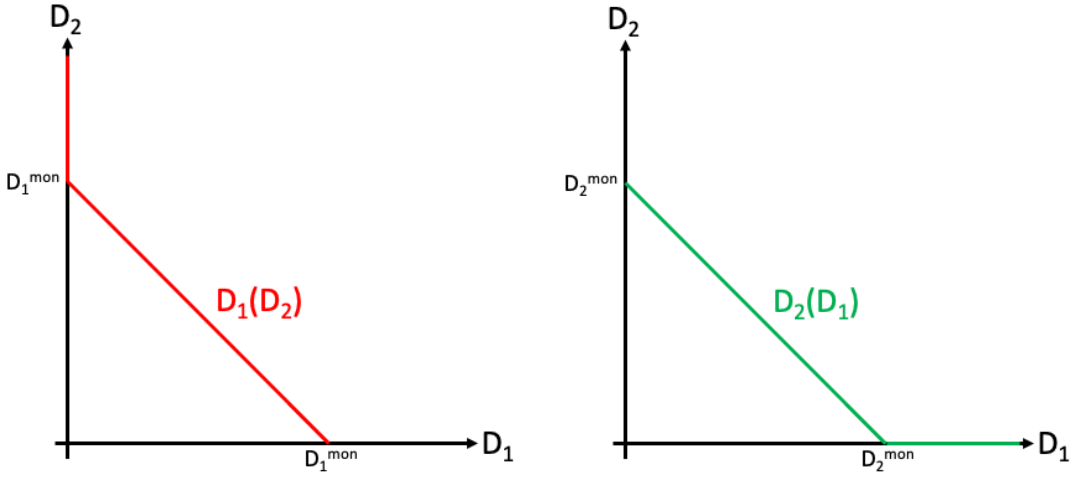
$$\frac{A}{\frac{q_2}{c} - s} D_2 = N \int_0^{\infty} \int_{p+\beta t}^{\infty} \int_0^{y} 2\pi r D_2 \exp\left(-\pi r^2 (D_1 + D_2)\right) f(\beta, V) \, \mathrm{d}r \, \mathrm{d}V \, \mathrm{d}\beta \quad \text{(H.4)}$$

where $y = \frac{(V - \beta t - p)v}{\beta}$ represents the maximum distance to the closest idle vehicle that a traveler with reservation value V and value of time β is willing to accept to prefer either platform to her outside option (see Appendix G).

Note that $D_1 = 0$ is always a solution to Equation H.3, and it is the only one for high values of D_2. At low values of D_2, there is a second solution with positive D_1. I assume that in this range, this positive solution will be the one reached in a potential equilibrium. In this range, the relationship between D_1 and D_2 is linear with

a slope of -1. Reductions in D_2 cause an equivalent increase in D_1. When $D_2 = 0$, D_1 equals the density of idle vehicles that a monopolist platform would reach setting prices (p, q_1) (D_1^{mon}).[2] The implicit function $D_1(D_2)$ determined by Equation H.3 then looks as in Figure H.1a. Similarly, Figure H.1b depicts the implicit function $D_2(D_1)$ determined by Equation H.4.

Figure H.1: Density of idle vehicles on one platform as a function of the density on the other.



(a) D_1 as a function of D_2. (b) D_2 as a function of D_1.

Notes: The function $D_1(D_2)$ is implicitly defined by Equation H.3. The function $D_2(D_1)$ is implicitly defined by Equation H.4. D_1^{mon} and D_2^{mon} denote the densities of idle vehicles that would be reached by a monopolist platform setting prices (p, q_1) and (p, q_2) respectively.

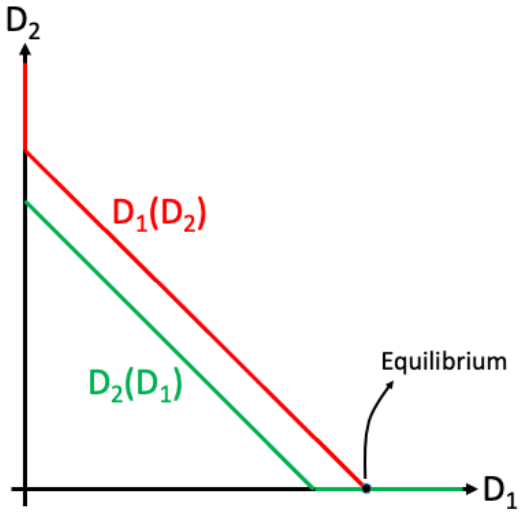
If $q_1 = q_2$ (equal pricing by platforms on both sides of the market), the graphs of these two implicit functions coincide on their linear and positive portions. A multitude of equilibria then exist under symmetric pricing, ranging between the extremes in which only one platform wins the entire market. It would be reasonable to assume, however, that symmetric pricing leads to the symmetric equilibrium in which $D_1 = D_2$ (platforms get equal market shares).

Increases in the price paid to drivers shift the respective graphs outward from the

[2] I assume $D_1^{mon} > 0$, which implies that the pair of prices (p, q_1) is sensible in that it allows a monopolistic equilibrium with positive numbers of riders and drivers (p not too high and/or q_1 not too small, see Appendix D). Note that sensible pricing always implies $q > cs$. Otherwise, drivers can never reach their reservation wage.

origin (e.g. increases in q_1 shift the graph of $D_1(D_2)$ outward). If $q_1 > q_2$, the graphs then look as in Figure H.2. Platform 1 wins the entire market. From a position of symmetric pricing ($p_1 = p_2, q_1 = q_2$), which may lead to equal market shares, either platform can slightly increase the price per trip paid to drivers in order to capture the entire market.

Figure H.2: One platform wins the entire market.



Notes: This figure shows the equilibrium densities of idle vehicles (D_1, D_2) when platforms charge the same price per trip to riders ($p_1 = p_2$) but Platform 1 pays a higher price per trip to drivers ($q_1 > q_2$).

References

Akbar, P., Couture, V., Duranton, G., & Storeygard, A. (2020, October). *Mobility and congestion in world cities: Evidence from google maps.* (Presented at the NBER Conference Economics of Transportation in the 21st Century)

Akbar, P. A., & Duranton, G. (2017). *Measuring the cost of congestion in a highly congested city: Bogotá* (Working paper No. 2017/04). CAF - Development Bank of Latin America.

Alvarez, F. E., & Argente, D. O. (2020). *On the effects of the availability of means of payments: The case of uber.* (NBER Working Paper Series)

Arnott, R. (1996). Taxi travel should be subsidized. *Journal of Urban Economics*, *40*, 316-333.

Azuara, O., Gonzalez, S., & Keller, L. (2019). *Who drives on ride-hailing platforms in latin america?* (Technical Note No. IDB-TN-1779). Inter-American Development Bank.

Balding, M., Whinery, T., Leshner, E., & Womeldorff, E. (2019). *Estimated TNC share of VMT in six US metropolitan regions* (Tech. Rep.). Fehr and Peers.

Basso, L. J., & Zhang, A. (2007). Congestible facility rivalry in vertical structures. *Journal of Urban Economics*, *61*, 218-237.

Benhabib, J., & Bisin, A. (2018). Skewed wealth distributions: Theory and empirics. *Journal of Economic Literature*, *56*(4), 1261-1291.

Bian, B. (2018, April). *Search frictions, netwrok effects and spatial competition: Taxis versus uber.*

Bimpikis, K., Candogan, O., & Saban, D. (2019). Spatial pricing in ride-sharing networks. *Operations Research*, *67*(3), 744-769.

Boffa, F., Fedele, A., & Iozzi, A. (2020, May). *Congestion and incentives in the age of driverless cars* (Research Paper No. 484). CEIS Tor Vergata.

Borenstein, S., Bushnell, J., & Stoft, S. (2000). The competitive effects of transmission capacity in a deregulated electricity industry. *The RAND Journal of Economics*, *31*(2), 294-325.

Brueckner, J. K. (2002). Airport congestion when carriers have market power. *American Economic Review*, *92*(5), 1357-1375.

Buchanan, J. M. (1956). Private ownership and common usage: The road case re-examined. *Southern Economic Journal*, *22*(3), 305-316.

Buchanan, J. M. (1969). External diseconomies, corrective taxes, and market structure. *American Economic Review*, *59*(1), 174-177.

Buchholz, N. (2020, February). *Spatial equilibrium, search frictios and dynamic efficiency in the taxi industry.*

Button, K. (2020). The transition from pigou's ideas on road pricing to their application. *Journal of the History of Economic Thought*, *42*(3), 417-438.

Castillo, J. C. (2019, December). *Who benefits from surge pricing?*

Castillo, J. C., Knoepfle, D., & Weyl, E. G. (2018). *Surge pricing solves the wild goose chase.*

Chen, M. K., Chevalier, J. A., Rossi, P. E., & Oehlsen, E. (2019). The value of flexible work: Evidence from uber driers. *Journal of Political Economy*, *127*(6).

Cheung, S. N. S. (1973). The fable of the bees: An economic investigation. *The Journal of Law and Economics*, *16*(1), 11-33.

Cohen, P., Hahn, R., Hall, J., Levitt, S., & Metcalfe, R. (2016). *Using big data to estimate consumer surplus: The case of uber.* (NBER Working Paper Series)

Cramer, J., & Krueger, A. B. (2016). Disruptive change in the taxi business: The

case of uber. *American Economic Review: Papers and Proceedings*, *106*(5), 177-182.

Duranton, G., & Turner, M. A. (2011, October). The fundamental law of road congestion: Evidence from us cities. *American Economic Review*, *101*, 2616-2652.

Erhardt, G. D., Roy, S., Cooper, D., Sana, B., Chen, M., & Castiglione, J. (2019). Do transportation network companies decrease or increase congestion? *Science Advances*, *5*(5).

Fowlie, M., Reguant, M., & Ryan, S. P. (2016). Market-based emissions regulation and industry dynamics. *Journal of Political Economy*, *104*(1).

Frechette, G. R., Lizzeri, A., & Salz, T. (2019). Frictions in a competitive, regulated market: evidence from taxis. *American Economic Review*, *109*(8), 2954-2992.

Greenstein, S., Peitz, M., & Valletti, T. (2016). Net neutrality: A fast lane to understanding the trade-offs. *Journal of Economic Perspectives*, *30*(2), 127-150.

Hall, J. D., Palsson, C., & Price, J. (2018). Is uber a substitute or complement for public transit? *Journal of Public Economics*, *108*, 36-50.

Hall, J. V., Horton, J. J., & Knoepfle, D. T. (2020, June). *Ride-sharing markets re-equilibrate.* (Unpublished)

Hall, J. V., & Krueger, A. B. (2018). An analysis of the labor market for uber's driver-partners in the united states. *Industrial and Labor Relations Review*, *71*(3), 705-732.

Heintzelman, M. D., Salant, S. W., & Schott, S. (2009). Putting free-riding to work: A partnership solution to the common-property problem. *Journal of Environmental Economics and Management*, *57*, 309-320.

Inrix. (2019). *Inrix global traffic scorecard.* Retrieved from https://inrix.com/scorecard/

Knight, F. H. (1924). Some fallacies in the interpretation of social cost. *The Quarterly Journal of Economics*, *38*(4), 582-606.

Lam, C. T., & Liu, M. (2017). *Demand and consumer surplus in the on-demand economy: the case of ride sharing* (Woring paper). MIT Initiative on the Digital Economy.

Leccese, M. (2021). *Disruptive competition and the cost of leveling the playing field: Evidence from the taxi industry.* (Available at SSRN)

Li, S., Tavafoghi, H., Poolla, K., & Varaiya, P. (2019). Regulating TNCs: Should Uber and Lyft set their own rules? *Transportation Research Part B*, *129*, 193-225.

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory.* Oxford University Press.

Mills, D. E. (1981). Ownership arrangements and congestion-prone facilities. *American Economic Review*, *71*(3), 493-502.

Mohring, H. (1972). Optimization and scale economies in urban bus transportation. *American Economic Review*, *62*(4), 591-604.

New York City Taxi and Limousine Commission. (2019). *New york state's congestion surcharge.* Retrieved from https://www1.nyc.gov/site/tlc/about/congestion-surcharge.page

Oviedo, D., Granada, I., & Perez-Jaramillo, D. (2020). Ridesourcing and travel demand: Potential effects of transportation network companies in bogotá. *Sustainability*, *12*(5).

Parry, I. W. H., & Small, K. A. (2005). Does Britain or the United States have the right gasoline tax? *American Economic Review*, *95*(4), 1276-1289.

Parry, I. W. H., & Small, K. A. (2009). Should urban transit subsidies be reduced? *American Economic Review*, *99*(3), 700-724.

Pigou, A. C. (1920). *The economics of welfare* (First ed.). London: Macmillan.

Salant, S., & Seegert, N. (2018). Should congestion tolls be set by the government or by the private sector? the knight-pigou debate revisited. *Economica*, *85*, 428-448.

San Francisco County Transportation Authority. (2017). *TNCs Today* (Tech. Rep.). Retrieved from https://www.sfcta.org/emerging-mobility/tncs-today

Secretaría Distrital de Movilidad, Bogotá D.C. (2019). *Encuesta de movilidad 2019.* Retrieved from https://www.simur.gov.co/portal-simur/datos-del-sector/encuestas-de-movilidad/

Shapiro, M. H. (2018, May). *Density of demand and the benefit of uber.* (Job market paper)

Silva, H. E., & Verhoef, E. T. (2013). Optimal pricing of flights and passengers at congested airports and the efficiency of atomistic charges. *Journal of Public Economics*, *106*, 1-13.

Small, K. A. (2012). Valuation of travel time. *Economics of Transportation*, *1*, 2-14.

Small, K. A., & Verhoef, E. T. (2007). *The economics of urban transportation.* Routledge.

Smichowski, B. C. (2018). Is ride-hailing doomed to monopoly? theory and evidence from the main u.s. markets. *Revue d'economie industrielle*, *162*, 43-72.

Spence, A. M. (1975). Monopoly, quality, and regulation. *Bell Journal of Economics*, *6*(2), 417-429.

Statista. (2020). *Market share of the leading ride-hailing companies in the united states.* Retrieved from https://www.statista.com/statistics/910704/market-share-of-rideshare-companies-united-states/

Tan, H., & Wright, J. (2018). A price theory of multi-sided platforms: Comment. *American Economic Review*, *108*(9), 2758-2760.

Tirachini, A. (2020). Ride-hailing, travel behavior and sustainable mobility: an international review. *Transportation*, *47*(4), 2011-2047.

Verhoef, E. T., Nijkamp, P., & Rietveld, P. (1997, September). The social feasibility of road pricing: A case study for the randstad area. *Journal of Transport Economics and Policy*, *31*(3), 255-276.

Verhoef, E. T., & Small, K. A. (2004). Product differentiation on roads: Constrained congestion pricing with heterogeneous users. *Jorunal of Transport Economics and Policy*, *38*(1), 127-156.

Weyl, E. G. (2010). A price theory of multi-sided platforms. *American Economic Review*, *100*, 1642-1672.

Yanocha, D., & Mason, J. (2019). *Ride Fair: A policy framework for managing Transportation Network Companies* (Tech. Rep.). Institute for Transportation and Development Policy.