THE INDUSTRIAL ORGANIZATION OF TRANSPORT MARKETS:
MODELING PRICING, INVESTMENT AND REGULATION IN RAIL AND ROAD NETWORKS
This thesis is part of a European Research Council project on ‘Optimizing Policies for Transport: Accounting for Industrial Organization in Network Markets’ (OPTION), funded under the European Union’s Seventh Framework Programme (FP7/2007-2013) through Advanced Grant №246969. The ERC’s financial support is gratefully acknowledged.

ISBN 978 90 361 0430 2

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul.

This book is no. 612 of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.
THE INDUSTRIAL ORGANIZATION OF TRANSPORT MARKETS: MODELING PRICING, INVESTMENT AND REGULATION IN RAIL AND ROAD NETWORKS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Economische Wetenschappen en Bedrijfskunde
op woensdag 8 april 2015 om 11.45 uur
in het auditorium van de universiteit,
De Boelelaan 1105

door

Adriaan Hendrik van der Weijde

geboren te Woerden
promotor: prof.dr. E.T. Verhoef
copromotor: dr. V.A.C. van den Berg
“...In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast Map was Useless, and not without some Pitilessness was it, that they delivered it up to the Inclemencies of Sun and Winters. In the Deserts of the West, still today, there are Tattered Ruins of that Map, inhabited by Animals and Beggars; in all the Land there is no other Relic of the Disciplines of Geography.”

# Contents

Preface

1 Introduction
   1.1 Transport markets ............................................. 1
   1.2 The industrial organization of transport markets ............. 2
   1.3 Modeling pricing, investment and regulation in rail and road networks ............................................. 3
   1.4 Structure of this book ......................................... 8

2 Price differentiation and discrimination in transport networks ........ 11
   2.1 Introduction .................................................. 11
   2.2 Differentiation and discrimination on a single link ........... 15
   2.3 Network effects - serial links ................................ 24
   2.4 Network effects - parallel links ............................ 27
   2.5 Conclusions .................................................. 30
   2.6 Appendix .................................................... 31

3 Competition in multi-modal transport networks: a dynamic approach 35
   3.1 Introduction .................................................. 35
   3.2 Methodology .................................................. 37
   3.3 Static model .................................................. 39
   3.4 Dynamic model ................................................ 41
   3.5 Numerical analysis ........................................... 49
   3.6 Conclusions .................................................. 55

4 A Hotelling model with price-sensitive demand and asymmetric distance costs: the case of strategic transport scheduling 57
   4.1 Introduction .................................................. 57
   4.2 Methodology .................................................. 59
   4.3 Social optimum ............................................... 60
   4.4 Full market separation ....................................... 60
   4.5 Equilibria with covered markets ............................ 62
   4.6 Other games .................................................. 66
   4.7 Discussion ................................................... 70
   4.8 Conclusions .................................................. 71

5 Stochastic user equilibrium traffic assignment with price-sensitive demand: do methods matter (much?) 73
   5.1 Introduction .................................................. 73
   5.2 Theory ......................................................... 75
   5.3 Simulation methodology .................................... 82
   5.4 Simulation results ........................................... 85
Contents

5.5 Conclusions .................................................................................. 92

6 Modeling the formation of transport networks and its regulation .......... 95
  6.1 Introduction .................................................................................. 95
  6.2 Modeling methodology .................................................................. 97
  6.3 Theory ......................................................................................... 104
  6.4 Numerical simulations ................................................................... 106
  6.5 Conclusions ................................................................................ 111

7 Conclusions ...................................................................................... 115
  7.1 Results and implications ............................................................... 115
  7.2 Avenues for future research .......................................................... 118

Bibliography ......................................................................................... 121

Samenvatting (Dutch summary) ............................................................. 129
Preface

The thesis you are about to read (or about to put back on the shelf after reading this preface) has been four years in the making. Much has happened in those four years, and many people deserve to be thanked for their roles in the coming together of this thesis, and in my life. Before that, however, I want to thank those without whom I would never have started writing this thesis at all. I was lucky enough to attend RA (now UCR) during the first few years of its existence. My lecturers and tutors there not only taught me what I needed to know to get into a good Master’s programme, but also showed me how academic research can be fun. Nevertheless, after my postgraduate studies, I had enough of being a student. I decided to get a job instead of pursuing a doctoral degree, and managed to find work at the University of Cambridge. I was hired to provide research assistance to a visiting researcher from the US, but rather than treating me as an assistant, he let me do my own research, introduced me to lots of interesting people, let me present my work at conferences, and in general showed me, again, how much fun academia is. Within six months, I was googling PhD programmes. Many thanks for that, Ben; I hope to pay it forward by giving my own students and staff the same opportunities you gave me. Many thanks also to all others at the EPRG for all the support I received.

Having decided I wanted to get a PhD, the two body problem started to kick in. Fortunately, I found Erik Verhoef, who offered me a position at the same university my wife had already been accepted, although he would not have know that at the time. Erik and Vincent, my second supervisor, turned out to be an excellent supervisory team, letting me do my own thing where possible, and subtly challenging, correcting or changing the direction of what I was doing when necessary. This thesis could not have been completed without them, and I am very grateful for all their help. I am also grateful to my other colleagues at the VU, particularly those who I shared an office with for all or most of the four years I was there. Alexandros, Hugo, Jan, Maria, Ruben; thank you for the discussions about research, discussions about anything but research, beer tastings, dinners, and everything else. It would not have been fun without you. Thanks also to the RE secretaries, Elfie and Jenny, who were always ready to help.

The final months of a PhD programme are always difficult, with lots of writing and formatting to do, forms to submit, and deadlines to remember. I am grateful to the Institute for Energy Systems at the University of Edinburgh for promising me a job at the end of it, which was an excellent incentive to wrap everything up as soon as possible and gave me something to look forward to. Of course, when a draft version of a thesis is complete, there are still some hurdles to take. I would therefore like to thank my committee: Adriaan Soetevent, André de Palma, Bruno De Borger, Jos van Ommeren and Serge Hoogendoorn, for the time they have invested in my thesis and their helpful comments and suggestions. Thanks also to my paranymphs, Martin and Hugo (who agreed to be my paranymph despite having to defend his own thesis immediately after me!), for their support.

PhD candidates in The Netherlands are employees of the university, which means that, at least on paper, they have a fixed number of working hours and a good work-life balance. That does not always work well in real life, but it did for me, to a large extent because of
family and friends, particularly all those at the Ruysdaelstraat, who gave me a reason to get away from work and do completely different things. But, of course, the most heartfelt thanks are reserved for my wife, Jeanine. ‘For better, for worse’—we have certainly experienced both in the past four years, and it was only because of you that the former far exceeded the latter. I hope and trust that we will continue to live happily ever after!

Finally, thanks to you, reader! You have made it to the end of this preface; I cannot realistically ask for more, but hope there will at least be some things in what follows that will arouse your interest in what I think is a very interesting field that will continue to be important for the foreseeable future. If you were hoping to see your name mentioned here but did not, that is only because of my forgetfulness, for which I apologize.

Edinburgh, February 2015
1 Introduction

1.1 Transport markets

The field of economics is concerned with the allocation of scarce resources, such as time, money, goods, or services. Often, these scarce resources are allocated in markets: places, physical or not, where buyers and sellers meet. For economists, markets are particularly interesting if they fail to deliver the best possible outcomes. ‘Best possible’, in this context, can mean multiple things: usually, a Pareto optimum (in which no single market participant can be made better off without harming another) or a social optimum (society as a whole cannot do better). In either case, any market outcome below the optimum gives regulators a reason to intervene; the question is then, of course, which interventions can restore the efficiency of the market in the best way.

Markets can fail to deliver Pareto- or socially optimal outcomes for various reasons, collectively called ‘market failures’. In transport markets, the most common market failure is the presence of externalities: costs or benefits of, in this case, consumption, that are not accrued by the market participant that decides how much to consume, but by others, who do not have a direct influence on those decisions. Many transport markets have negative environmental externalities: carbon emissions, local air pollution, and noise are primarily incurred by people other than those who drive, fly, or transport goods. Most transport markets also suffer from congestion externalities: an additional car on the road slows down all other drivers (Pigou, 1920; Knight, 1924; Vickrey, 1969). Both externalities result in sub-optimal market outcomes, and hence, large costs to society; consequently, they have been researched extensively. Everything else equal, they increase consumption above its (social) optimum: if car drivers were forced to pay for the decrease in travel time of all other road users, they would travel less, or at different moments; if air passengers had to pay for their share of the plane’s carbon emissions, fewer people would fly, and so on. Usually, these externalities are analyzed on their own, assuming that there are no other market failures. For instance, studies often assume that markets are perfectly competitive (e.g., there are so many airlines that each airline cannot influence ticket prices on its own, but takes these prices as given), that there is one monopolist who faces perfectly elastic demand (such that, again, it cannot influence prices), or assume price-taking behavior in some other way.

However, in practice, market participants do often exhibit non-price taking behavior: there are monopolists who face imperfectly elastic demand and oligopolists that are large enough to influence prices on their own. This market failure, in many settings, cannot be assumed away. As private investment in, and ownership of, transport networks is increasing, it will only become more important. This thesis therefore aims to explore the industrial organization of transport markets: that is, the decisions made by non-price taking market participants, the outcomes of those decisions, and the ways in which regulators can influence them. It develops methods to model these decisions, and uses those to investigate pricing and scheduling behavior, investment, and regulation. In contrast to much of the existing literature, it does so while explicitly taking the networked nature of transport markets into account.
1 Introduction

1.2 The industrial organization of transport markets

This thesis analyzes the industrial organization of transport network markets. As such, it builds on the existing industrial organization literature, generalizing existing theories, applying them to transportation markets, and developing methodological advancements. Cournot (1838) was one of the first to develop mathematical models in which the behavior of suppliers with market power could be analyzed. Modeling an oligopolistic market where several competitors supply the same product, and choose the quantities they produce to maximize their private profits, Cournot shows how, in this setting, prices are lower than in a monopoly, but still higher than the suppliers’ marginal costs. Only if the number of suppliers approaches infinity do prices approach marginal costs, and hence, the social optimum. Each supplier sets its own quantity, but takes into account how this influences the decisions of other market participants, since these, in turn, affect market prices; in equilibrium, no supplier has an incentive to unilaterally change its produced quantity.\(^1\)

Bertrand (1883), in a review of Cournot’s work, proposed a different model, in which competitors do not set quantities, but prices. If this is the case, prices always equal marginal costs if there is more than one supplier; private profits are always equal to zero if marginal costs are constant. Both models still form the basis for most analyses of oligopolistic markets, although they have been extended in all possible directions. One important direction, taken by, amongst others, Hotelling (1929) and Salop (1979), extends oligopolistic theories to include product differentiation. They show how, if suppliers can differentiate their products, they can avoid head-on competition, and still make positive profits even if they compete on prices. This differentiation can also be spatial in nature; suppliers can be located in different places. Hotelling and Salop’s models only include one dimension; recently, these models have been extended to more general networks (e.g. Heijnen and Soetevent, 2014). One of the chapters of this thesis will apply a generalized version of Hotelling’s model to transport scheduling: there, products are differentiated in time.

In many cases, and most definitely in transport markets, products can not only be substitutes, but also complements, such that consumers derive more value from a combination of products. Travelers may, for instance, travel to a railway station by bus, and then continue their journey by train: they need both modes to reach their destination. Economides and Salop (1992) study a situation where substitutes and complements are produced by competing suppliers. As one may expect, competition in the presence of substitutes decreases prices, and increases social welfare. However, if competing operators offer complements, this is not the case. Rather, as each competitor disregards the negative effect of a price increase on its competitor’s profits, all operators set prices that are higher than the prices a monopolist would set. This effect is called ‘double marginalization’, since more than one competitor exerts market power. Hence, although increasing the number of competitors may be beneficial from a societal viewpoint if they offer substitutes, this is not the case for complements.

These results are directly relevant to transport markets. In a transport network, links and modes can be substitutes for each other, but they may also be complementary. Often, however, the distinction may not be very clear: links are typically neither pure substitutes nor pure complements for all travelers carried. There are other differences between transport markets and other network markets. Not only does transport usually have external effects; demand- and supply structures are also often very complex. Travelers are not just consumers,

\(^1\)Hence, in his book, Cournot also already develops the equilibrium concept that would, much later, be known as the pure strategy Nash equilibrium.
1.3 Modeling pricing, investment and regulation in rail and road networks

but also supply some of the inputs for travel; most importantly, they supply their time. This has important implications for optimal prices, investment, and regulation (see e.g., Mohring, 1972).

In the past decade, a separate literature on the industrial organization of transport networks has therefore emerged. Usually, the models developed in these studies have a very simple network representation. de Palma and Lindsey (2000) consider a network with one origin and one destination, connected by multiple roads. They analyze several forms of ownership, and show how important ownership is. Private ownership by competing operators can, for instance, increase welfare relative to a situation in which a private operator competes with publicly owned roads. Brueckner (2002), in the context of air transport, also illustrates the importance of market power. If airlines are large enough, the congestion externalities in the market should not simply be internalized by charging each airline a toll equal to its marginal external costs (a classic result in the road pricing literature), since each airline already internalizes the congestion it imposes on its own other flights. The toll should thus reflect external costs imposed on other airlines’ flights.

De Borger et al. (2005) also look at parallel competition, but differentiate between long-distance and short-distance travelers, in a setting where each link is regulated by a separate government, which primarily cares about its own residents. De Borger et al. (2008) develop a similar model that also includes serial links. Pels and Verhoef (2007) model a three-node network, in which there is both parallel and serial competition. They consider pricing, capacity choices and regulation, and conclude that optimal regulation is highly dependent on the competitive structure of the market.

Research on larger networks is scarce; although there is empirical work on the modeling of traffic flows, some of which will be discussed below, pricing, investment, and regulation are not generally discussed. One exception is Adler et al. (2010), who analyze competition between high-speed rail and air travel in Europe, using a large network model.

The present thesis adds to this developing literature, extending and combining it.

1.3 Modeling pricing, investment and regulation in rail and road networks

This is not an empirical thesis; instead, it develops theoretical models: simplified mathematical representations of transport markets, which can be used to develop theories about markets that do not yet exist, or explore which effects one would expect to find in real-world markets. Every model makes assumptions. Even if it would be possible to create a model that captured a real-world market down to the smallest detail, that model would be of limited use: because of its complexity it would be difficult to derive new insights from it, and neither would it be able to predict what would happen to the market if something changed. Naturally, some details are essential, and cannot be assumed away. The restrictiveness of the assumptions, and hence, the level of detail in a model, is, as it should be, always a subject of discussion. In the transportation community, engineers generally prefer models that are more realistic, as these can be used to provide specific answers to questions about real-world networks. Economists, conversely, prefer simpler models, with more restrictive assumptions, as their outputs are easier to interpret, which can lead to more insight into the mechanisms at play.

Although several chapters draw heavily on the transportation engineering literature, this book is primarily an economic thesis. It therefore focuses more on insight than on detailed
1 Introduction

descriptions of reality, and to do so, it makes restrictive assumptions. Each chapter focuses on a different setting, and hence, each model is more detailed than others in at least one dimension. When modeling transport markets, several classes of assumptions are particularly important, and it is worth looking at those before moving on to specific models.

Assumptions about the physical network.

As already mentioned, transport markets rarely exist in isolation; they are usually part of larger networks, which consist of several nodes, markets for travel between them, and links that connect them. When building a model of a particular transport market, one must choose how much of this network should be included. Specifically, one must choose how large the network model will be, what the network configuration will look like (in particular, how great the degree of substitutability and complementarity will be, i.e., will the model have parallel and/or serial links, substitute modes, etc.?), and whether, or how, the dynamic nature of transportation networks will be considered.

In the transport economics literature, network models are usually small, to help derive insight. Many papers focus on networks with just two nodes, and one or two links. Chapters 2 and 4 of this thesis also use such models. The addition of more complexity would make these models less tractable, without necessarily giving more insight. Chapter 3 develops a slightly larger model with two modes and three nodes, and thus four links in total: the minimum number required to analyze parallel and serial competition together. Chapters 5 and 6, on the other hand, analyze a much larger network, since this is necessary to analyze the effects of overlapping routes on user equilibria, and the spatial allocation of investment.

If the size of the network is important, the network configuration is even more crucial. As mentioned above, the presence of substitute links or modes, as well as complementary links or modes, is one of the most important determinants of the social desirability of competition. With the exception of chapter 2, which discusses a monopoly, all models developed in this thesis therefore use networks that have both complementary and substitute links. In chapter 3, the substitute links are defined as separate modes (rail and road); in the other chapters, modes are not explicitly defined. The larger networks in chapters 5 and 6 use meshed grid networks, precisely because this type of network, by definition, includes a large number of complementary and substitute links, of roughly equal sizes. It is worth noting that, in the air transport literature, networks that are larger than one of two links are often of the ‘hub and spoke’-variety: central nodes (the hubs) are connected to the other nodes, while these other nodes are not connected to each other (see, e.g., Brueckner, 2004; Silva et al., 2014). Although passenger rail and road networks sometimes also include hubs and spokes, these networks are usually more meshed. Pure hub and spoke networks are therefore not considered in this thesis.

As the title of this thesis suggests, it deals primarily with rail and road networks. What sets these modes apart from others, and most importantly from air travel, is that there are high costs associated with the construction of new links and with the destruction of existing links. Naturally, there are also important differences between rail and road travel; in most countries, for instance, infrastructure ownership and operation are decoupled in rail networks, but not in road networks. Because of the nature of this thesis, many of these differences will be ignored; indeed, in most chapters, the modes will not be explicitly defined.
Assumptions about user costs.

Transportation users face several different types of costs. Of particular relevance here are the monetary cost (the fare, road price, or petrol cost), the time cost, the crowding cost, and the cost of arriving at a different time than travelers prefer (the schedule delay cost). Monetary costs are included in all models developed in this thesis, but always represent only road prices or fares; petrol costs and other operational and maintenance costs are never included explicitly. This simplification is unlikely to affect the results, as these costs are often rather constant and hardly vary with pricing or capacity decisions.

The time costs of travel are also important in all chapters. All types of travel take time, which usually comes at a cost to users. Using an exogenous value of time (VOT)\(^2\), these costs can be monetized. Naturally, VOTs may differ across travelers, which can have important implications (see, e.g., Arnott et al., 1988; van den Berg and Verhoef, 2011). This thesis will disregard this complication, and instead assumes that all users have the same VOT. They may, however, differ in other respects, as we will discuss below.

For most public transport modes, travel times are fixed: they do not, or not to a large degree, depend on the number of users. For roads, the opposite is true. Road congestion has a large impact on travel times: in the US alone, congestion has been estimated to cause an additional 5.5 billion hours of travel time in 2011, amounting to a loss of $121 billion when combined with the cost of wasted fuel (Schrank et al., 2012). It is therefore important to include the relation between traffic flows and travel speeds in theoretical models. Various relationships have been proposed in the literature. Static models, which do not have a time dimension, usually assume that travel times are a continuous function of travel flows (see e.g., Vickrey, 1963); the static models in this thesis will make the same assumptions. Some chapters use linear functions for simplicity; others, for road links, use the so-called BPR function (US Bureau of Public Roads, 1964), which has been shown to be a good approximation of average long-distance road travel costs.

There is more variety in the treatment of congestion in dynamic models. Many transport studies use the bottleneck model (Vickrey, 1969; Arnott et al., 1990), which assumes that transport links have a fixed capacity; if travel flows exceed this capacity, a queue forms, which dissipates only when flows drop below the link capacity. This model is very useful in many settings, but is problematic in larger network models. It is discontinuous (congestion only has an effect if flows are above link capacities), and serial bottlenecks are difficult to analyze (if two bottlenecks in series have different capacities, only one of them will be relevant). An alternative approach, proposed first by Lighthill and Whitham (1956) and Richards (1956), and therefore often referred to as the ‘LWR’ model, instead uses flow-concentration curves, which are commonly used in kinematics, to relate changes in flows to changes in speeds. Chapter 3 of this thesis uses a simplified version of this model, as proposed by Henderson (1974) and Chu (1995), which assumes that each individual user travels at a constant speed and shockwaves travel at the same speed as the traffic that carries them. More detailed classes of models exist (e.g., car-following models, which model the behavior of individual drivers. See May (1990) for an example), but these are too complex to be used in the analyses presented in this thesis.

Although public transport travel times may be fixed, travelers using these modes may experience crowding costs (see, e.g., Wardman and Whelan, 2011; Li and Hensher, 2011). Crowding may occur on the platform, when entering the vehicle, and during travel in the

---

\(^2\)Some (mostly empirical) studies prefer the term ‘value of travel time savings’ (VTTS)
1 Introduction

vehicle. This thesis only considers the last, and arguably the most important category: in-vehicle crowding. For simplicity, crowding costs are assumed to be linear; the difference between sitting down and standing is ignored.

Finally, users may face schedule delay costs; that is, costs that are incurred because users do not arrive at their preferred arrival time, but earlier or later. These costs can be substantial (Small, 1982). In the chapters that deal with static models, which do not include arrival times, schedule delay costs are, naturally, not included: they do play an important role in the other chapters. These assume that users have a fixed unit cost of schedule delay (i.e., incur a cost for every minute they are late or early), such that the total schedule delay costs are linear in the amount of schedule delay. This assumption has been made in most of the existing literature. Chapter 3 assumes that the costs of being early and late are, per unit of time, the same; chapter 4 specifically focuses on a situation in which the costs of being late are higher than the costs of being early.

There are, of course, other types of user costs associated with transportation that are not considered in this thesis. Travel times are, for instance, uncertain, and this uncertainty is costly. This cost of travel time variation, or travel time uncertainty is currently receiving more attention in the literature (see, e.g., Fosgerau and Karlstrom, 2010; Koster et al., 2011), with very interesting results. However, uncertain travel times are beyond the scope of this thesis, and hence, these costs will not play any role in the analyses presented below.

Assumptions about the behavior of individual users, and the formation of user equilibria.

Specifying user costs is one thing; determining how they influence travel behavior is another. Two assumptions are important here. The first determines how users decide which of several different alternatives (e.g., routes though a network, modes, or departure times) they will use. The most traditional approach (Wardrop, 1952) is to assume that, in equilibrium, all alternatives that are used must have the same user cost, which is lower than the cost of all unused routes. In the resulting Wardropian, or deterministic, equilibrium, no single user can lower its costs by choosing another alternative. This concept is used in chapters 2–4. Chapters 5 and 6 instead use stochastic user equilibrium (SUE) models, based on random utility theory (McFadden, 1978). These assume that, in addition to the deterministic user costs, each user’s costs also include a stochastic error term, which is unobservable, but comes from a known distribution. Using that distribution, one can calculate the probability that a given user prefers a certain alternative over another. These probabilities, summed across all users, can then be used to calculate expected traffic flows. Naturally, if the variance of the error term approaches zero, the SUE collapses to the deterministic user equilibrium. Stochastic user equilibrium models are useful; they arguably describe reality better, and can be more computationally convenient. There are several types of SUEs, each with its own assumptions about the distributions of the error terms; chapters 5 and 6 explain this in more detail.

The second assumption deals with the price-sensitivity of demand. In chapter 3, the total demand for travel is assumed to be fixed, to isolate the effects of departure time decisions. In the other chapters, demand does vary with travel costs, either in a very general form (chapter 2), linearly (chapter 4), or in a more complex fashion (chapters 5–6, where the models include an alternative to not travel, which is treated in the same way as the physical routes in the description of choice behavior).
1.3 Modeling pricing, investment and regulation in rail and road networks

Assumptions about operator costs, and the behavior of operators.

It costs money to maintain roads and railways, and to operate trains, and these costs have been well-researched. Since this thesis focuses on competition between operators, it generally assumes that these operation and maintenance costs are negligible, or that the marginal costs of transporting an extra user are constant, such that they do not influence the operators’ pricing rules. This is, obviously, a simplification: in the real world, operators have different marginal costs. However, the inclusion of this complication in oligopolistic models has disadvantages too, as differences in marginal costs can obscure other effects, and make drawing conclusions about other competitive mechanisms difficult.

There are many ways to model competition in markets. Markets may be perfectly competitive, such that each market participant takes market prices as given; this situation may sometimes be a good benchmark, but real-world markets, and especially transport markets, are rarely perfectly competitive. Perfect competition is therefore not the focus of this thesis. The opposite situation, in which there is only one monopolist, is common in transport markets, and it is therefore discussed in this thesis, particularly in chapter 2. In between these two extremes, oligopolistic markets have a (usually small) number of market participants; chapters 3–4 and 6 analyze this situation. As explained above, there are different forms of oligopolistic competition. First of all, operators may compete on prices (Bertrand competition), or on quantities (Cournot competition). In road- and rail markets, arguably, Bertrand competition better describes the market; quantities follow directly from travel decisions taken by users, in response to prices set by operators. In air travel, conversely, Cournot competition may be better, and is indeed used often used in theoretical studies. Secondly, there are different equilibrium concepts. This thesis, like most of the literature, will mostly consider Nash equilibria: equilibria in which no market participant can do better by unilaterally changing its decisions. Chapter 4 also considers Stackelberg equilibria, in which operators make decisions sequentially, as may happen in some real-world settings. There are other, more exotic equilibrium concepts, such as strong Nash equilibria and coalition-proof Nash equilibria. These assume that there are possibilities for collusion between market participants, which is not generally allowed, and therefore not taken into account here.

Assumptions about the regulatory environment

Few markets are completely unregulated; transport markets are no exception. All chapters in this thesis therefore address regulation. Throughout, it is assumed that regulators maximize social welfare, that there is only one such a regulator, and that market participants have no choice but to obey it. In the real world, this is not always true: regulation takes place on different levels and may differ across jurisdictions, regulators may not always maximize social welfare, and there are different levels of enforcement. Indeed, there is a substantial literature on tax competition between authorities (e.g., De Borger et al., 2005), the presence of multiple regulators (e.g., Pels and Verhoef, 2004), and related topics; this is outside the scope of this thesis. As usual in the transport literature, social welfare is purely utilitarian: income effects are ignored. This is justifiable, as transportation costs do not usually constitute large parts of consumers’ budgets.

Apart from social welfare maximization, real-world regulators might also be interested in the acceptability and equity of their policies. In a perfect world, these would issues would not change much, as the social welfare maximum is Kaldor-Hicks efficient: if some groups are unhappy with the policies but society as a whole benefits, the winners can always be made
to compensate the losers, such that everybody is better off. For this reason, the economic models used in this thesis do not address acceptability or equity. Naturally, in models that are more detailed representations of particular real-world settings, or that are used to directly inform policy, they need to be considered, as redistribution may not always be possible.

The different chapters model different types of regulation. First-best regulation, in which the regulator chooses all relevant variables to maximize social welfare, without any constraints, is usually the benchmark. Second-best regulation, in which the regulator is somehow constrained, because it cannot directly control some variables or is restricted in its choice of values (e.g., because it can only levy positive taxes, and not give subsidies), but maximizes welfare subject to these constraints, is also considered in most chapters. Chapter 6 also addresses quasi-first-best regulation, in which the regulator cannot control some variables, but ignores this, and uses the first-best rules for the remaining instruments.

1.4 Structure of this book

This introductory chapter is followed by five technical chapters and a conclusion. As discussed above, all models of transport markets can be categorized according to various characteristics, related to assumptions about demand, costs, the behavior of market participants and regulators, and the physical characteristics of the market. For the purposes of this thesis, it is useful to focus on the distinction between static models, which do not have a time dimension, and dynamic models, which do. Moreover, the models developed in the various chapters have different time horizons. Chapters 2 and 3 examine pricing strategies, assuming that capacities are given. Chapter 4 also considers scheduling decisions. Chapter 5 develops methods that can be used to analyze both pricing and longer-term decisions, such as investment, while chapter 6 focuses on investment decisions. Table 1.1 shows how all technical chapters fit into this framework.

Naturally, other distinctions can be made, but Table 1.1 does show that the technical chapters, together, span a wide modeling space. All chapters propose ways to model and explain the effects of the industrial organization of transport markets on prices, welfare, and optimal regulatory strategies, but each chapter focuses on a different aspect or setting. Hence, they can also be read separately. All chapters have a dual purpose: they introduce novel methods, or show how existing methods can be applied to answer questions related to the industrial organization of transport markets, and then also apply these methods to gain new insight.

Chapter 2 analyzes the effects of price differentiation and discrimination by a monopolistic transport operator, which sets fares in a congestible network. Using three models, with different spatial structures, it describes the operator’s optimal strategies in an unregulated market, a market where price differentiation is not allowed (i.e., ticket prices must be the same for all users), and a market where price discrimination is illegal (i.e., ticket prices must only differ with the marginal external costs of users), and analyze the welfare effects of uniform and non-discriminatory pricing policies. The three models allow for separate consideration of the three different forms of price differentiation and discrimination in networks: by user class, by origin-destination pair, and by route. This chapter generalizes the existing literature on price discrimination, in which groups usually only differ in their value of time to also include differences in marginal external costs. In this setting, non-differentiated and non-discriminatory policies may increase or decrease welfare, and non-discrimination can be worse than non-differentiation. The results obtained for a single-link network can be general-
1.4 Structure of this book

<table>
<thead>
<tr>
<th>Static models</th>
<th>Dynamic models</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. A Hotelling model with price-sensitive demand and asymmetric distance costs: the case of strategic transport scheduling</td>
<td></td>
</tr>
<tr>
<td>5. Stochastic user equilibrium traffic assignment with price-sensitive demand: do methods matter (much)?</td>
<td></td>
</tr>
<tr>
<td>6. Modeling the formation of transport networks and its regulation</td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1: Structure of chapters 2–6

ized to a situation where operators price-discriminate or differentiate based on users’ origins and destinations, but not directly to a situation in which differentiation is based on the routes users take.

Chapter 3 moves to networks in which transport operators compete. It analyzes the behavior of market participants in a multi-modal commuter network, where roads are not priced, but public transport has a usage fee, which is set while taking the effects on the roads into account. In particular, it analyzes the difference between markets with a monopolistic public transport operator, which operates all public transport links, and markets in which each public transport link is owned by a separate operator. Importantly, users not only choose which mode to use, but also decide when to travel. In this dynamic setting, even if the total travel demand is inelastic, serial Bertrand-Nash competition on the public transport links leads to different fares than a serial monopoly. This results from the fact that trip timing decisions, and therefore the generalized prices of all commuters, are influenced by all fares in the network. The chapter ends with a numerical simulation, which shows that, contrary to the results obtained in classic studies on vertical competition, monopolistic fares are not always lower than duopolistic fares.

Like chapter 3, chapter 4 also considers departure time choices, but this time, with a focus on how these choices affect scheduling decisions made by transport operators. It proposes a generalized Hotelling horizontal differentiation model with price-sensitive demand and asymmetric distance costs. In this model, two competitors choose fares and departure times in a fixed time interval; consumers’ locations indicate their desired departure times. The model is used to show how departure times can be strategic instruments, and how they are best regulated.

Chapter 5 analyzes how user equilibrium assignment models impact outcomes, and which assignment models are best. It compares three competing stochastic user equilibrium traffic
1 Introduction

assignment methodologies (multinomial probit, nested logit, and generalized nested logit), using a congestible transport network. The models are used to evaluate policy decisions, such as profit-maximizing tolling or second-best socially optimal tolling. The results are then used to investigate how these optimal tolls, and their performance, depend on the model choice, and hence, how important the differences between models are. As it turns out, the differences between models are small, as a result of the congestibility of the network, as long as they are calibrated correctly. Hence, it may be better to use computationally more efficient logit models instead of probit models, in at least some applications, even if the latter are preferable from a conceptual viewpoint.

Chapter 6 uses the results from chapter 5 to analyze investment in networks; hence, in contrast to the earlier chapters, it focuses on long-term decisions. It proposes methods to analyze the effects of different policies that control the formation of transport networks by private operators. These are then used to analyze investment in a simple network with a limited number of nodes and routes, and possibilities to build congestible travel links with discrete capacities between each pair of nodes. In this setting, fare regulation and location-independent capacity subsidization are not always sufficient to achieve the first-best welfare-maximizing solution. Second-best fare regulation can be better than second-best network regulation in some, but not all, situations.

Following the technical chapters, chapter 7 concludes the thesis.
2 Price differentiation and discrimination in transport networks

2.1 Introduction

This chapter investigates the effects of allowing or disallowing transport operators to charge different users different tolls or fares. Charging different tolls or fares to different users is common practice in many transportation markets. This differentiation in prices is particularly interesting in the context of transportation, as, unlike in markets for most consumer goods, there usually consumption externalities associated with travel. Moreover, these externalities are usually asymmetric: the congestion externality imposed by, for instance, a passenger car driver on a truck driver using the same road at the same time is unlikely to be the same as the externality imposed by the truck on the car driver. There may also be other reasons, apart from these externalities, for different users to have different marginal costs: e.g., some types of vehicles cause more road damage than others, drivers of cars outfitted with transponders are easier to toll than drivers who pay in cash, and in air travel, the weight of passengers and their luggage directly affects fuel consumption. Travel products are also difficult to resell, and different groups of users can often be easily distinguished, which makes it easy to differentiate prices.

It is not surprising, therefore, that many existing studies look at these issues in the context of transportation markets, both empirically and theoretically. Adachi (2005) formulates a model with linear demand functions and consumption externalities, in which all users have the same marginal external costs, but different marginal private costs. In this linear model, monopolistic third-degree price differentiation does not change output, but can increase welfare if it improves the composition of users. Nocke et al. (2011) show how a monopolist can use advance-purchase discounts to price-discriminate. Closest to this chapter, Czerny and Zhang (2015), in the context of air travel, use a model with general demand- and cost functions, in which different users have different values of time, but the same marginal external cost. They conclude that ticket-price discrimination can increase welfare, even if it reduces the total number of travelers. If all users have the same value of time, and their inverse demand- and cost curves are linear, price discrimination does not change this total number of travelers.

This chapter aims to make three contributions to this literature. First of all, we will consider a situation in which users do not only have different values of time (and thus, different marginal costs), but also different marginal external costs. This difference can, for instance, arise if vehicles have different sizes (e.g. trucks and cars), and thus, impose different congestion externalities on other users. This is well-established in the empirical literature (see e.g., Al-Kaisy et al., 2002), but not usually incorporated in studies on price discrimination.

Secondly, these differences in marginal external costs also allow us to distinguish between...
two situations in which different users are charged different fares or tolls. Rather than calling all instances where different users or user groups are charged different prices for the same product ‘price discrimination’, we will distinguish between price discrimination and price differentiation, and examine both pricing regimes separately. If \( I \) is the set of all user classes (or individual users if each class only has one member) and \( f_i \) the toll or fare charged to a particular class \( i \), we say that price differentiation occurs if

\[
\exists \{i, j\} \in I^2 \quad f_i \neq f_j
\]  

(2.1)

If, in addition \( c_i \) is the cost of travel (not including tolls) faced by user class \( i \) and \( n_i \) the number of users from class \( i \) that travel, and assuming that the marginal costs of transport provision are independent of the user class, price discrimination occurs when

\[
\exists \{i, j\} \in I^2 \quad f_i - \sum_k (\partial c_k / \partial n_i) n_k \neq f_j - \sum_k (\partial c_k / \partial n_j) n_k
\]  

(2.2)

In other words, price differentiation occurs if different user groups are charged different tolls or fares, and price discrimination occurs if groups are charged different tolls based only on their demand structures, rather than costs. This definition of price discrimination (Eq. 2.2) is common in the microeconomics and industrial organization literature\(^2\) (see, e.g., Stigler, 1986; Tirole, 1988; Stole, 2007; Verboven, 2008); in the more recent transportation literature (e.g. Czerny and Zhang, 2015), on the other hand ‘price discrimination’ is often defined as in Eq. 2.1: what we call ‘differentiation’.\(^3\) If there are no external costs associated with consumption, or if the external costs are the same regardless of which consumer buys the product, the two definitions are the same. In transportation markets, however, users often have different external costs (e.g., consider the example of truck- and car drivers). Hence, price differentiation can occur without discrimination; indeed, disallowing differentiation can lead to discrimination. We therefore think that, in our context, it is particularly important to make a distinction between the two. Price differentiation occurs often, and can easily be justified on the basis of ‘fairness’: trucks, for instance, are often charged higher tolls than passenger cars. Price discrimination, on the other hand, is often perceived as ‘unfair’, since it differentiates between users only on the basis of their willingness to pay, not on a difference in costs. Nevertheless, it is also practiced in transportation markets, e.g., in the form of discounts for users above a certain age.

The third contribution of this chapter is that we explicitly consider the networked nature of transportation markets. Most studies look at one market in isolation. This allows for the analysis of discrimination between users or user classes, and often produces tractable, interpretable results. For this reason, we too will start our analysis with a simple single-market model, in which several user classes travel on one link, and the operator discriminates and/or differentiates between these classes. In a network, however, users from different origins travel to different destinations. When deciding what to charge for use of a link, an operator can then also differentiate and/or discriminate based on users’ origin-destination (OD) pairs. If there are multiple links connecting two nodes, an operator could also differentiate by route. These last two types of differentiation and discrimination can only be analyzed using a network model.

\(^2\)There, it is usually a difference in marginal production costs that leads to differentiation without discrimination, but the idea is the same.

\(^3\)Our terminology is not new: see, for example, Chen and Schwartz (2013)
A model that combines all these forms of price discrimination and differentiation would be far less tractable, so we examine three separate models. In the first (Fig. 2.1), one congestible link is shared by two classes of users, each with its own inverse demand function. If the monopolistic operator is not allowed to differentiate between the two classes, it sets one fare $f$, which is paid by all users. If it can differentiate, it sets two fares, $f_1$ and $f_2$, which are paid only by the users from the first and second class, respectively. If differentiation is allowed, but discrimination is not, these two fares can differ only as much as the marginal costs of the two user groups differ. In transport markets, this type of price differentiation may be achieved by, for instance, charging cars and trucks different road prices.

In the second model (Fig. 2.2), we include serial network effects, and consider price discrimination and differentiation by OD-pair. There are now three nodes, two links in series connecting them, and three groups of users: two groups short-distance travelers traveling from node 1 to node 2 and from node 2 to 3, and long-distance travelers traveling from 1 to 3; each has its own inverse demand function. Congestion functions may differ between links, and long-distance travelers incur congestion costs on both links. Hence, while in the first model two groups of users impose congestion externalities on each other, this model has three groups, each of which imposes a congestion externality on some, but not all, others. If the operator can differentiate between OD-pairs, it sets three fares, one for each class of users. If the long-distance travelers pay more than the sum of the two short-distance fares, there is differentiation on at least one of the two links. If the operator cannot differentiate, all users pay the same for use of each link, and hence, long-distance fares are equal to the sum of the two short-distance fares. Again, if differentiation is allowed, but discrimination is not, the operator still sets three fares, but any difference between the long-distance fare and the sum of the two short-distance fares has to be related to marginal cost differences. This type of differentiation is similar to what is sometimes called ‘price discrimination by bundling’ (Adams and Yellen, 1976; Zhang and Czerny, 2012). Although it is common in public transport markets (the ticket price for travel from A to C via B is often different from the sum of the prices for travel from B to C and B to C), it is certainly not used everywhere; road prices, for instance, are often additive.

In the third model (Fig. 2.3), we look at the effects of parallelity in networks. There is only one class of users, but two parallel links, and hence, two routes, with different congestion functions. Here, we consider price differentiation by route. In contrast to the previous two
models, users only impose congestion externalities on other users taking the same route; however, the choices they make still affect all users through the inverse demand function. If the operator can differentiate between routes, it sets two fares $f_1$ and $f_2$, to be paid by $N_1$ users taking the first route, and $N_2$ users taking the second, respectively. If it cannot differentiate, it sets one fare $f$, which is paid by all users, regardless of their route choice. If it can differentiate but not discriminate, $f_1$ and $f_2$ differ only with the marginal cost functions of the two user classes. This type of price differentiation (or its absence) is also present in many public transport markets: one ticket for travel between two cities may, or may not be, valid on a number of routes.

Together, these three models encompass the range of options for price differentiation and discrimination in public transport markets. They are, to some extent, variations on the same theme, and could be incorporated in one, generalized network model. Treating them separately, however, allows us to consider the different effects of price differentiation and discrimination in isolation. In each of the models, we derive the first-best socially optimal fares, and three sets of private fares: 1) as they would be set by an unrestricted monopolist, 2) in a situation where the monopolist is not allowed to differentiate, and 3) in a situation where the monopolist is allowed to differentiate, but cannot discriminate. We then compare these outcomes, to see if and when price differentiation and/or discrimination can improve welfare. As far as possible, we use general cost- and demand functions; in some cases, we need to impose linearity to produce meaningful results.

Naturally, we can not consider all possible forms of differentiation and discrimination in this chapter. Specifically, in line with most of the existing literature, we limit the scope of our analysis in four ways. First, we restrict ourselves to third-degree discrimination (sometimes called ‘group discrimination’) and differentiation; implying that different groups or classes of users with different characteristics are charged different tolls. This is not the only form

\[ f_1 \] and \[ f_2 \] differ only with the marginal cost functions of the two user classes. This type of price differentiation (or its absence) is also present in many public transport markets: one ticket for travel between two cities may, or may not be, valid on a number of routes.

Together, these three models encompass the range of options for price differentiation and discrimination in public transport markets. They are, to some extent, variations on the same theme, and could be incorporated in one, generalized network model. Treating them separately, however, allows us to consider the different effects of price differentiation and discrimination in isolation. In each of the models, we derive the first-best socially optimal fares, and three sets of private fares: 1) as they would be set by an unrestricted monopolist, 2) in a situation where the monopolist is not allowed to differentiate, and 3) in a situation where the monopolist is allowed to differentiate, but cannot discriminate. We then compare these outcomes, to see if and when price differentiation and/or discrimination can improve welfare. As far as possible, we use general cost- and demand functions; in some cases, we need to impose linearity to produce meaningful results.

Naturally, we can not consider all possible forms of differentiation and discrimination in this chapter. Specifically, in line with most of the existing literature, we limit the scope of our analysis in four ways. First, we restrict ourselves to third-degree discrimination (sometimes called ‘group discrimination’) and differentiation; implying that different groups or classes of users with different characteristics are charged different tolls. This is not the only form

\[ f_1 \] and \[ f_2 \] differ only with the marginal cost functions of the two user classes. This type of price differentiation (or its absence) is also present in many public transport markets: one ticket for travel between two cities may, or may not be, valid on a number of routes.

Together, these three models encompass the range of options for price differentiation and discrimination in public transport markets. They are, to some extent, variations on the same theme, and could be incorporated in one, generalized network model. Treating them separately, however, allows us to consider the different effects of price differentiation and discrimination in isolation. In each of the models, we derive the first-best socially optimal fares, and three sets of private fares: 1) as they would be set by an unrestricted monopolist, 2) in a situation where the monopolist is not allowed to differentiate, and 3) in a situation where the monopolist is allowed to differentiate, but cannot discriminate. We then compare these outcomes, to see if and when price differentiation and/or discrimination can improve welfare. As far as possible, we use general cost- and demand functions; in some cases, we need to impose linearity to produce meaningful results.

Naturally, we can not consider all possible forms of differentiation and discrimination in this chapter. Specifically, in line with most of the existing literature, we limit the scope of our analysis in four ways. First, we restrict ourselves to third-degree discrimination (sometimes called ‘group discrimination’) and differentiation; implying that different groups or classes of users with different characteristics are charged different tolls. This is not the only form

\[ f_1 \] and \[ f_2 \] differ only with the marginal cost functions of the two user classes. This type of price differentiation (or its absence) is also present in many public transport markets: one ticket for travel between two cities may, or may not be, valid on a number of routes.

Together, these three models encompass the range of options for price differentiation and discrimination in public transport markets. They are, to some extent, variations on the same theme, and could be incorporated in one, generalized network model. Treating them separately, however, allows us to consider the different effects of price differentiation and discrimination in isolation. In each of the models, we derive the first-best socially optimal fares, and three sets of private fares: 1) as they would be set by an unrestricted monopolist, 2) in a situation where the monopolist is not allowed to differentiate, and 3) in a situation where the monopolist is allowed to differentiate, but cannot discriminate. We then compare these outcomes, to see if and when price differentiation and/or discrimination can improve welfare. As far as possible, we use general cost- and demand functions; in some cases, we need to impose linearity to produce meaningful results.

Naturally, we can not consider all possible forms of differentiation and discrimination in this chapter. Specifically, in line with most of the existing literature, we limit the scope of our analysis in four ways. First, we restrict ourselves to third-degree discrimination (sometimes called ‘group discrimination’) and differentiation; implying that different groups or classes of users with different characteristics are charged different tolls. This is not the only form
2.2 Differentiation and discrimination on a single link

of discrimination. Under first-degree price discrimination, or differentiation, each individual consumer can be charged a different price. Although this normally improves welfare, as it allows the producer to capture the whole consumer surplus in addition to its producer surplus, it is not usually allowed or even possible. Second-degree price discrimination, in which consumers are charged a price based on the quantity they consume, is much more common, but less so in passenger transportation markets.

Second, note that, in the transportation literature, a distinction is often made between ‘full prices’, i.e., private costs plus fares or tolls, and ‘ticket prices’, i.e., the fares or tolls only. Hence, there are also two potential types of differentiation and discrimination: in full prices and in ticket prices. If private costs depend on total usage levels, as they usually do in transportation markets, these two types are not the same. As the definitions in Eqs. 2.1–2.2 already indicates, this chapter only considers ticket price discrimination and differentiation. Although others have also looked at discrimination in full prices (e.g. Czerny and Zhang, 2015), this type of discrimination is not usually observed in real-world markets, and would unnecessarily complicate the analysis.

Third, we only consider internal solutions: i.e., outcomes in which at least some users from each class travel. Corner solutions, in which one group does not travel, are certainly interesting, but difficult to analyze in a general setting, and the inclusion of the various corner solutions that exist would overcomplicate our exposition. We compare and contrast four cases: the social optimum, a monopoly, a monopoly where price differentiation is not allowed, and a monopoly where price discrimination is not allowed. We do not consider other forms of competition (e.g. oligopolies). We also do not comment on the impacts of marginal changes in price differences or discrimination on welfare, except where that is necessary to analyze a monopoly.

This chapter shows that restricting a monopolist to charge uniform or non-discriminatory prices may increase or decrease social welfare, depending on the parameters of the model. Moreover, non-discriminatory pricing is not necessarily better than uniform pricing, even though the former is arguably the most ‘fair’ policy. In contrast to the simpler models developed in the existing literature, here, both policies have an impact on the total number of users even if all users have the same value of time. The results obtained from a single-link model can be generalized to a situation with serial links, where discrimination or differentiation is based on the origins and destinations of users. They cannot be generalized directly to a setting with parallel links, where differentiation is based on the routes that users take.

2.2 Differentiation and discrimination on a single link

2.2.1 Social optimum

In this model, there are two user classes, each with its own inverse demand function \( D_1(N_1) \) and \( D_2(N_2) \). Both inverse demand functions are continuously differentiable with first-order derivatives \( D_1'(N_1) < 0 \) and \( D_2'(N_2) < 0 \). The two classes share the same link, and hence, impose negative congestion externalities on each other. Since the two classes may be composed of different types of vehicles, these externalities need not be symmetric; adding an additional user of one class may have a much larger effect on congestion than adding an additional user of the other. Apart from this difference in external costs, the two classes may also have different marginal private cost functions. They may, for instance, have a different value of time; although subject to the same level of congestion, this would lead the
two classes to face different user costs. Reflecting these two potential differences, we define two user cost functions $c_1 = \alpha c (\beta N_1 + N_2)$, and $c_2 = c (\beta N_1 + N_2)$, where $\beta$ is a relative congestion coefficient, which captures the marginal external costs of class 1 users relative to this in class 2 (e.g., a passenger car equivalent). The $\alpha$ parameter measures how class 1 users value congestion relative to class 2 (e.g., a relative value of time).

We assume that $c(\cdot)$ is continuously differentiable with $c'(\cdot) > 0$. Social welfare is then the sum of the integrals of both demand functions minus the total user costs:

$$W = \int_0^{N_1} D_1(n) \, dn + \int_0^{N_2} D_2(n) \, dn - (\alpha N_1 + N_2) c(\beta N_1 + N_2)$$  \hfill (2.3)

Depending on the functional forms of the inverse demand and average user cost functions, it may be optimal to have only one group traveling. Assuming that both groups travel, a social planner sets fares $f_1$ and $f_2$ (charged to the first and second user class, respectively) and flows $N_1$ and $N_2$ to maximize $W$ s.t.

$$D_1(N_1) - \alpha c (\beta N_1 + N_2) - f_1 = 0$$  \hfill (2.4)

$$D_2(N_2) - c (\beta N_1 + N_2) - f_2 = 0$$  \hfill (2.5)

where the constraints ensure that marginal user costs plus fares equal marginal benefits, and hence, that the resulting equilibrium is consistent with the users’ preferences.

Maximizing $W$ subject to these constraints gives $f_1 = (\alpha N_1 + N_2) \beta c'$ and $f_2 = (\alpha N_1 + N_2) c'$, where, $c' = \partial c / \partial n_2$ (i.e. if users in class 2 have passenger cars, the increase in cost resulting from the addition of one extra passenger-car equivalent); fares are equal to marginal costs, such that all external costs are internalized by the users (see also Pigou, 1920; Knight, 1924). Following these pricing rules leads to differentiation if $\beta \neq 1$ as, in that case, users have different marginal costs; there is no discrimination: any differences in fares between the two groups are related to marginal cost differences.

### 2.2.2 Unrestricted monopoly

A monopolist maximizes $\pi = f_1 N_1 + f_2 N_2$ subject to the same constraints as the social planner (Eqs. 2.4–2.5). The resulting fares are:

$$f_1 = (\alpha N_1 + N_2) \beta c' - N_1 D_1'$$  \hfill (2.6)

$$f_2 = (\alpha N_1 + N_2) c' - N_2 D_2'$$  \hfill (2.7)

where $c' = \partial c / \partial n_2$ and $D_i' = \partial D_i / \partial n_i$. Again, this is not surprising. The monopolist charges a markup, in addition to the user’s marginal external costs: both prices are higher than socially optimal, while both $N_1$ and $N_2$ are lower. Importantly, there is price differentiation if $\beta \neq 1$ and/or $D_1(n) \neq D_2(n)$. Price discrimination only occurs if $D_1(n) \neq D_2(n)$, regardless of the value of $\beta$.

---

6Passenger car equivalents (PCEs), first introduced in the 1965 Highway Capacity Manual (HRB, 1965) are widely used in road transportation studies.

7E.g., both external and private costs of class 1 users are a linear function of those of class 2. In general, this need not be the case, but without this assumption our models would become intractable.
2.2 Differentiation and discrimination on a single link

Comparing welfare under unrestricted monopolistic pricing to the more restrictive policies without differentiation or discrimination that we will examine below is difficult without strong assumptions on the inverse demand- and cost functions. We can, however, analyze in which situations the composition of users is optimal, and how the total number of users changes with these policies.

Given a total number of users \( N_1 + N_2 \), the composition of users (or share of users from each class) is optimal if \( \Delta A \equiv \partial W/\partial N_1 - \partial W/\partial N_2 = 0 \). As long as this equality holds, it is not possible to increase welfare by decreasing the number of users from one class while simultaneously increasing the number of users of the other class by the same amount. Using 2.3,

\[
\Delta A = (D_1 - \alpha c) - (D_2 - c) + (\alpha N_1 + N_2)(1 - \beta) c' (2.8)
\]

where the first two terms give the marginal change in private user benefits resulting from the marginal composition change, and the third term gives the marginal change in total costs. Substituting the first-order conditions of the unrestricted monopoly in this expression, this can be simplified to \( \Delta A = N_2 D_2' - N_1 D_1' \). Unless \( N_2 D_2' = N_1 D_1' \) (which is unlikely to happen unless the two demand functions are equal), \( \Delta A \neq 0 \); the composition of users is not optimal in an unrestricted monopoly. Whether \( \Delta A \) is positive or negative (and hence, whether \( N_1 \) or \( N_2 \) is higher then optimal) depends on the relative cost- and demand functions of the two classes. In theory then, a more restrictive pricing policy, which disallows price discrimination or differentiation, can improve welfare, even if it does not increase (or even reduces) the total number of travelers. This effect of a policy on the composition of users is sometimes called the ‘allocation effect’ (see, e.g., Czerny and Zhang, 2015).

Besides this allocation effect, a policy change may also affect the total number of travelers. This may be called the ‘output effect’. Although it is theoretically possible to analyze this output effect in a similar way as the allocation effect (i.e. by evaluating \( (N_1/ (N_1 + N_2)) \partial W/\partial N_1 + (N_2/ (N_1 + N_2)) \partial W/\partial N_2 \): keeping the composition of users constant, but increasing the total), this is not useful here, as the total usage level will never be optimal in any of the monopolistic settings we examine. Using a metric like this, we could therefore only conclude that welfare can, in all cases, be increased by increasing the number of travelers, but not how it changes between two monopolistic settings. We can, however, examine how the restrictive policies change total usage levels, as we will see below.

This approach, combining information about the effect of restrictive policies on output with an analysis of the allocative efficiency of the resulting equilibria, is useful, because it does not need any assumptions about demand or congestion. It does, however, also have disadvantages. We can only say that a policy unambiguously increases welfare if it leads to an equilibrium where \( \Delta A = 0 \), and does not decrease the total usage level. If \( \Delta A \neq 0 \), or the total number of users changes, we can only describe the output- and allocation effects, without knowing how they combine to affect aggregate welfare. Therefore, we will complement our analytical approach with a numerical example, in which we can more explicitly show how welfare changes when differentiation or discrimination is now allowed.

2.2.3 Monopoly without price differentiation

This situation is similar to the above, except that \( f_1 = f_2 = f \). Hence, the monopolist now maximizes \( \pi = f (N_1 + N_2) \) s.t.

\[
D_1 (N_1) - \alpha c (\beta N_1 + N_2) - f = 0 \quad (2.9)
\]
Price differentiation and discrimination in transport networks

\[ D_2 (N_2) - c (\beta N_1 + N_2) - f = 0 \]  \hspace{1cm} (2.10)

The optimal monopolistic fare can, in this case, be written as

\[
\begin{align*}
f &= (N_1 + N_2) c' \left( \frac{D_1' + \alpha \beta D_2'}{D_1' + D_2' - (\alpha - 1)(\beta - 1)c'} \right) \\
- N_1 D_1' &\left( \frac{D_2'}{D_1' + D_2' - (\alpha - 1)(\beta - 1)c'} \right) \\
- N_2 D_2' &\left( \frac{D_1'}{D_1' + D_2' - (\alpha - 1)(\beta - 1)c'} \right)
\end{align*}
\]  \hspace{1cm} (2.11)

This immediately shows why it is important to distinguish between price discrimination and differentiation: although there is now no differentiation, there is still discrimination if \( \beta \neq 1 \).

More importantly, enforcement of uniform pricing usually leads to price discrimination: if \( D_1 (n) = D_2 (n) \) (and hence, an unrestricted monopolist would not price discriminate, but only differentiate), the uniform price in Eq. 2.11 is discriminatory.

In general, the single fare \( f \) consists of two parts. The first internalizes (some of) the marginal external costs. Since these differ across the two classes, the monopolist uses a weighted average, with the weights determined by the relative slopes of the inverse demand curves, corrected for the fact that the second group has a different congestion parameter \( \beta \) and value of time \( \alpha \) than the first. These parameters are taken into account because they co-determine the sensitivity of each class to sub-optimal prices. The second term is a weighted average of the two monopolistic markups, where, again, the weights are determined by the slopes of the inverse demand curves, and, if \( \alpha \neq 1 \) or \( \beta \neq 1 \), the slope of the cost function. Each class’ weight is inversely proportional to its relative inverse demand slope; the more price-sensitive a class is (i.e., the flatter its inverse demand curve is), the closer the joint markup lies to what this class’ own markup would be in an unrestricted monopoly.

The weights also include the slope of the cost function such that they correctly reflect the sensitivity of each class’ demand, taking the effect of changes in congestion costs into account.

Eq. 2.11 also illustrates that it is important to consider differences in external costs and internal cost functions together; only if both are present does the cost function influence the markup. In that case, the weights on \( N_1 D_1' \) and \( N_2 D_2' \) do not add up to one: the monopolistic markup is not just somewhere between the two markups in an unrestricted setting. This happens because, if users differ in two dimensions, it is not possible to charge the correct average external cost and average markup; each average needs to take into account that the other average is distortionary.

Under a uniform pricing policy, \( \Delta A = (\alpha N_1 + N_2) (1 - \beta) c' \). Hence, if \( \beta = 1 \) (all users have the same marginal external cost; e.g., all drive the same type of vehicle), uniform pricing always leads to an optimal composition of users – a clear improvement over an unregulated monopoly. If \( \beta < 1 \), \( N_2 \) is larger than optimal; if \( \beta > 1 \), \( N_1 \) is larger than optimal. In these cases, uniform pricing could improve or deteriorate the composition of users, depending on the model parameters.

Determining the effect of uniform pricing on the total number of users is more complicated, as this total number is determined by the user equilibrium constraints, and does not have a closed form solution. Assuming, without loss of generality, that \( f_2 \geq f_1 \) and following Czerny and Zhang (2015), we can define direct demand functions \( N (f_1, f_2) \equiv N_1 (f_1, f_2) + N_2 (f_1, f_2) \).
2.2 Differentiation and discrimination on a single link

and a price difference \( \phi \equiv f_2 - f_1 \geq 0 \). The total derivative of \( N \) with respect to the price difference can then be written as:

\[
\frac{dN}{d\phi} = \frac{\partial N}{\partial f_1} \frac{df_1}{d\phi} + \frac{\partial N}{\partial f_2} \frac{df_2}{d\phi} \tag{2.12}
\]

\[
= \left( \frac{\partial N}{\partial f_1} + \frac{\partial N}{\partial f_2} \right) \frac{df_1}{d\phi} + \frac{\partial N}{\partial f_2} \tag{2.13}
\]

Totally differentiating the user equilibrium conditions and using Cramer’s Rule gives:

\[
\frac{\partial N_1}{\partial f_1} = (D'_2 - c') / \Omega < 0 \tag{2.14}
\]

\[
\frac{\partial N_2}{\partial f_1} = \beta c' / \Omega > 0 \tag{2.15}
\]

\[
\frac{\partial N_1}{\partial f_2} = \alpha c' / \Omega > 0 \tag{2.16}
\]

\[
\frac{\partial N_2}{\partial f_2} = (D'_1 - \alpha \beta c') / \Omega < 0 \tag{2.17}
\]

where \( \Omega = \begin{vmatrix} D'_1 - \alpha \beta c' & -\alpha c' \\ -\beta c' & D'_2 - c' \end{vmatrix} = (D'_1 - \alpha \beta c') (D'_2 - c') - \alpha \beta c'^2 > 0 \). All partials and cross-partials have the expected signs. Hence,

\[
\frac{\partial N}{\partial f_1} = (D'_2 - (1 - \beta) c') / \Omega \tag{2.18}
\]

\[
\frac{\partial N}{\partial f_2} = (D'_1 - \alpha (\beta - 1) c') / \Omega \tag{2.19}
\]

Interestingly, it is theoretically possible that one (but not both) of these partial derivatives is positive if \( \beta \) is either very small, or very large. Because the link is congested, a fare increase for one user class will increase the number of users from the other. In some cases, this increase may be larger than the decrease in users from the first class.

Finally, totally differentiating the operator’s profit function with respect to \( f_1 \) and \( \phi \) gives

\[
\frac{df_1}{d\phi} = -\frac{\partial^2 \pi}{\partial^2 \pi / (\partial f_1 \partial \phi)} \tag{2.20}
\]

where

\[
\frac{\partial^2 \pi}{(\partial f_1 \partial \phi)} = \frac{\partial N}{\partial f_2} + \frac{\partial N_2}{\partial f_1} + \frac{\partial N}{\partial f_2} + f_1 \left( \frac{\partial^2 N}{(\partial f_1 \partial f_2)^2} \right) \tag{2.21}
\]

\[
+ \phi \left( \frac{\partial^2 N_2}{(\partial f_1 \partial f_2)^2} + \frac{\partial^2 N}{(\partial f_2)^2} \right)
\]
Price differentiation and discrimination in transport networks

\[
\frac{\partial^2 \pi}{\partial f_1^2} = 2 \left( \frac{\partial N}{\partial f_1} + \frac{\partial N}{\partial f_2} \right) + f_1 \left( \frac{\partial^2 N_1}{\partial f_1^2} + 2 \frac{\partial^2 N_1}{\partial f_1 \partial f_2} + \frac{\partial^2 N_1}{\partial f_2^2} \right) + (f_1 + \phi) \left( \frac{\partial^2 N_2}{\partial f_1^2} + 2 \frac{\partial^2 N_2}{\partial f_1 \partial f_2} + \frac{\partial^2 N_2}{\partial f_2^2} \right) \tag{2.22}
\]

In a linear case, then,

\[
\frac{df_1}{d\phi} = \frac{(2D_1' + \beta (1 - \alpha) c' + \alpha (1 - \beta) c') / \Omega}{2 \left( \frac{\partial N}{\partial f_1} + \frac{\partial N}{\partial f_2} \right)} \tag{2.23}
\]

and hence

\[
\frac{dN}{d\phi} = \frac{(\alpha - \beta) c'}{2 \Omega} \tag{2.24}
\]

where \( \Omega \), in this linear case, is a positive constant. Since a uniform pricing policy reduces \( \phi \) from \( f_2 - f_1 \) to zero, it increases output by the negative of the integral of Eq. 2.24 which, in a linear case, is \( (f_2 - f_1) (\beta - \alpha) c' / 2\Omega \).

Contrary Czerny and Zhang (2015) (which, itself, generalizes Robinson, 1933), even if time valuations are the same for all users (\( \alpha = 1 \)), this is not enough for the output effect to be zero if cost- and demand functions are linear; this only holds if \( \alpha = \beta \) (such that each user’s relative value of time is equal to its relative external cost). This happens because, in their model, all users have the same marginal external cost. Since we have assumed that \( f_2 \geq f_1 \), and the increase in the number of users for a specific fare difference is given by the integral of Eq. 2.24, the number of users is higher under uniform pricing only if \( \alpha < \beta \); i.e., the user class that is charged the highest fare has a relatively low value of time, compared to its relative external cost.

For general demand- and cost functions, if \( \beta > 1 \) and \( df_1 / d\phi > 0 \) (i.e.; the lowest unrestricted fare is higher than the uniform fare), the total number of users always reduces if prices are differentiated (i.e.; increases if a uniform pricing policy is enforced). If \( df_1 / d\phi < -1 \) (i.e. the highest unrestricted fare is lower than the uniform fare), output always increases with price differentiation. Usually, though, one would expect \( -1 < df_1 / d\phi < 0 \), in which case the sign of the output effect is then determined by the values of \( \alpha, \beta \), and the Hessians of the two direct demand functions.

Taking the allocation and output effect together: it is, even in a linear world, only possible to determine that a uniform pricing policy is always welfare-enhancing if \( \beta = 1 \) (such that the allocation effect is definitely positive) and \( \alpha < 1 \) while \( f_2 \geq f_1 \) (such that the output effect is non-negative). In all other cases, a uniform pricing policy may or may not be welfare-enhancing, depending on the model parameters. In the numerical example below, we will explore this further.

2.2.4 Monopoly without price discrimination

Here, as in the unrestricted monopoly, the monopolist maximizes \( \pi = f_1 N_1 + f_2 N_2 \), now s.t. Eqs. 2.4–2.5 and

\[
f_1 - (\alpha N_1 + N_2) \beta c' = f_2 - (\alpha N_1 + N_2) c' \tag{2.25}
\]
2.2 Differentiation and discrimination on a single link

Assuming, for tractability, that users costs are linear, and thus, \( c_t = 0 \),

\[
f_1 = (\alpha N_1 + N_2) \beta c' - N_1 D_1' \frac{D_2' + (\alpha - \beta) c'}{D_1' + D_2' - 2\alpha (\beta - 1) c'} - N_2 (D_2' - (\beta - 1) c') \frac{D_1' + (\alpha + \beta - 2\alpha\beta) c'}{D_1' + D_2' - 2\alpha (\beta - 1) c'}
\]

\[
f_2 = (\alpha N_1 + N_2) c' - N_1 D_1' \frac{D_2' + (\alpha - \beta) c'}{D_1' + D_2' - 2\alpha (\beta - 1) c'} - N_2 (D_2' - (\beta - 1) c') \frac{D_1' + (\alpha + \beta - 2\alpha\beta) c'}{D_1' + D_2' - 2\alpha (\beta - 1) c'}
\]

In the non-differentiated setting above, the monopolist charged all users an average marginal cost, plus an average markup. Here, it charges each user class its actual marginal costs. This, in turn, also affects the markup. Even though this markup is, by definition, still independent of class, the operator exploits the fact that it can at least charge every class its own marginal costs, and consequently sets a higher markup than in Eq. 2.11. The weights on \( N_1 D_1' \) and \( N_2 D_2' \) are now considerably more complex, but are still inversely related to each class’ own inverse demand slope. If \( \beta \geq 1 \), such that the first user class pays the highest fare, the average markup moves closer to the markup charged to the first class in an unrestricted monopoly when the relative value of time \( \alpha \) of that class decreases.

The non-discrimination constraint ensures that \( \Delta A = 0 \): by definition, the marginal change in user benefits (i.e., the difference between \( f_1 \) and \( f_2 \)) resulting from a marginal change in user composition is equal to the marginal change in social costs. Hence, under this policy, the composition of users is always optimal and the allocation effect is always positive. Determining the effect of non-discriminatory pricing on output is more difficult, as Eq. 2.24 can not be applied here. In this case, we cannot simply define \( \phi \) as the difference in fares; \( \phi \) would then depend on \( N_1 \) and \( N_2 \), and cannot be treated as an exogenous variable. We must now define \( \phi = f_2 - f_1 - \Delta mec. \) where \( \Delta mec = (N_1 + \alpha N_2) (1 - \beta) c \). If discrimination is not allowed, \( \phi = 0 \). Using the same technique as before,

\[
\frac{dN}{d\phi} = \frac{\partial N}{\partial f_1} \frac{df_1}{d\phi} + \frac{\partial N}{\partial f_2} \frac{df_2}{d\phi}
\]

\[
= \left( \frac{\partial N}{\partial f_1} + \frac{\partial N}{\partial f_2} \right) \frac{df_1}{d\phi} + \frac{\partial N}{\partial f_2} \frac{df_2}{d\phi} + \frac{\partial N}{\partial f_2} \frac{d\Delta mec}{d\phi}
\]

Note the third term, which was not present in the no-differentiation analysis; this now appears because \( \phi \) is now not simply the difference in fares, but the difference in fares minus the difference in marginal external costs. This term can be written as

\[
\frac{d\Delta mec}{d\phi} = \left( \frac{dN_1}{d\phi} + \alpha \frac{dN_2}{d\phi} \right) (1 - \beta) c' + (N_1 + \alpha N_2) (1 - \beta) \frac{dc'}{d\phi}
\]

which only disappears if \( \beta = 1 \), and has no clear cut sign. The addition of this term means that, even if cost- and demand functions are linear, and \( \alpha = \beta \), the total usage level may be different under a non-discriminatory policy than under an unrestricted monopoly. As we will show in the numerical example below, it may be higher or lower, depending on
the model parameters. Hence, even though the absence or presence of price differentiation does not change the total number of users when $\alpha = \beta$, the absence or presence of price discrimination does have an effect.

In conclusion: a non-discriminatory policy always has a positive allocation effect: it always leads to an optimal composition of users. It may also increase the number of users, and hence, would be socially beneficial. In a linear world, this is guaranteed to happen if $\beta = 1$ (all users have the same marginal external cost) and the class that is charged the lowest fare also has the lowest value of time. However, like uniform pricing, non-discriminatory pricing may also decrease the number of users, and thus potentially lower welfare. Finally, there is no guarantee that non-discriminatory pricing is always better policy than uniform pricing, as we shall see in more detail below.

### 2.2.5 Numerical example

Fig. 2.4 illustrates the effects described above, for a very simple linear case. In all panels, $c = 10 + 0.1(\beta N_1 + N_2)$. In the panels on the left, (a) and (c), $D_1 = 100 - N_1$ and $D_2 = 65 - N_2$. In the other two panels, the inverse demand functions are reversed, such that $D_1 = 65 - N_1$ and $D_2 = 100 - N_2$. Finally, in the two top panels, (a) and (b), $\alpha = 1$, whereas in the bottom panels, $\alpha = 3$. The vertical axis in both panels denotes the relative efficiency $\omega$ of the non-differentiated and non-discriminatory monopolistic prices; i.e. the difference in welfare between these settings and an unrestricted monopoly, relative to the difference in welfare between the first-best social optimum and an unrestricted monopoly (Verhoef et al., 1995). If $\omega < 0$, an unrestricted monopoly is better than the restrictive policy; a higher $\omega$ means that a policy is closer to the social optimum.

Fig. 2.4 shows that, even with simple linear demand- and cost functions, the effects of uniform or non-discriminatory pricing policies are ambiguous. Either policy can potentially increase or decrease welfare, relative to an unrestricted monopoly, depending on the parameters of the model.

Looking at the results of uniform pricing: as we have shown above, the output effect disappears if $\alpha = \beta$ (at $\beta = 1$ in the top panels, and at $\beta = 3$ in the bottom panels). What is left over can only be the allocation effect. Clearly, this can be positive or negative. Hence, uniform pricing can improve welfare, but it can also decrease it. In this linear example, the latter is mostly likely to happen if the class with the highest reservation price also has a much higher external cost than the other class.

For non-discriminatory pricing, the output effect disappears only if $\alpha = \beta = 1$; in that case, only the positive allocation effect remains. In other cases, the output effect may be negative, and large enough to offset the allocation effect: it may also be positive, and work in the same direction as the allocation effect. Hence, enforcing non-discrimination sometimes, but certainly not always, improves welfare.

Uniform pricing is sometimes better than non-discriminatory pricing. In this linear example, this happens if the class with the highest reservation price has a lower external cost than the other class (e.g. a morning peak on a highway if the inverse demand function for cars always lies above the inverse demand function for trucks, as may be expected) while both classes have the same value of time. In that case, uniform pricing induces a monopolist to increase output, relative to an unrestricted setting, by so much that, even though a non-discriminatory policy would lead to a better composition of users, uniform pricing is still better. If the classes also have different values of time, this effect is parameter-dependent. This does show, however, that a pricing strategy that is considered the most ‘fair’ may not be
2.2 Differentiation and discrimination on a single link

\[ D_1 = 100 - N_1, D_2 = 65 - N_2, \alpha = 1 \]

\[ D_1 = 65 - N_1, D_2 = 100 - N_2, \alpha = 1 \]

\[ D_1 = 100 - N_1, D_2 = 65 - N_2, \alpha = 3 \]

\[ D_1 = 65 - N_1, D_2 = 100 - N_2, \alpha = 3 \]

Non-differentiation

Non-discrimination

Figure 2.4: Model 1 – Numerical example
the best in terms of social welfare. Naturally, if $\beta = 1$, uniform pricing and non-discrimination are the same. As panel (c) shows, there may also be other $\beta$'s for which this holds, even in this linear case.

2.3 Network effects - serial links

2.3.1 Social optimum

In the previous model, we considered a situation in which multiple user classes travel between the same two nodes, using one link. Here, we consider a setting with three nodes, connected by two serial links, as shown in Fig. 2.2. As before, there are different user classes; however, these differ not only in their values of time and congestion parameters, but also in origin and destination. Users from the first class travel between nodes 1 and 2, using only the first link. The second class uses the second link to travel between nodes 2 and 3, and, finally, the third class travels from 1 to 3, using both links.

We simplify this model by assuming that the two links have the same congestion functions $c$. We also assume that the inverse demand functions for short-distance travel (from 1 to 2 and from 2 to 3) are the same, such that the model becomes symmetric: fares and usage levels will be the same on both links. The short-distance inverse demand functions are then given by $D_s(N_s)$, where $N_s$ is the number of short-distance travelers on each of the two links, while the long-distance inverse demand function is given by $D_l(N_l)$. We do allow for the possibility of long-distance travelers having a different relative value of time $\alpha$, and a different congestion coefficient $\beta$, than short-distance travelers. Hence, all short-distance travelers face costs $c(\beta N_l + N_s)$, while long-distance travelers incur costs $2\alpha c(\beta N_l + N_s)$.

The assumed symmetry of this model allows us to obtain more compact expressions for the monopolistic fares, and for the welfare effects. However, as we will discuss below, and show in the appendix, the qualitative results do not depend on this assumption. A model with two different cost functions, one for each of the links, or different inverse demand functions for each of the three user classes, will not produce fundamentally different insights, especially when it comes to the effect of price differentiation on the total number of users.

Given a symmetric solution, social welfare is given by

$$W = 2 \int_0^{N_s} D_s(n) \, dn + \int_0^{N_l} D_l(n) \, dn - 2(\alpha N_l + N_s) c(\beta N_l + N_s)$$ (2.31)

A social planner sets the fares for long-distance and short-distance travel, $f_s$ and $f_l$, and the corresponding usage levels, maximizing $W$. Similar to the previous model, there are two user equilibrium constraints:

$$D_s(N_s) - c(\beta N_l + N_s) - f_s = 0$$ (2.32)
$$D_l(N_l) - 2\alpha c(\beta N_l + N_s) - f_l = 0$$ (2.33)

Maximizing $W$ subject to these constraints results in simple marginal-cost pricing rules: $f_s = (\alpha N_l + N_s) c'$ and $f_l = 2\beta (\alpha N_l + N_s) c'$. Even if all users have the same congestion coefficient $\beta$, long-distance travelers still pay a higher fare than short-distance travelers, simply because they cause congestion on two links instead of just one.
2.3.2 Unrestricted monopoly

If it can set separate fares for short- and long-distance travelers, the monopolist maximizes
\[ \pi = 2f_s N_s + f_l N_l \] s.t. Eqs. 2.32–2.33. With no restrictions on these fares, the monopolist
can differentiate and discriminate between users from different OD-pairs. If \( f_l \neq 2f_s \), long-
distance travelers pay a different fare for use of at least one of the two links than short-distance
travelers. This unrestricted setting is sometimes called ‘OD-based pricing’ (e.g., Ohazulike
et al., 2013), and leads to fares
\[ f_s = (\alpha N_l + N_s) \ell' - N_s D_s' \quad \text{and} \quad f_l = 2\beta (\alpha N_l + N_s) \ell' - N_l D_l'. \]
Again, these fares internalize external costs, and include a monopolistic demand-related
markup. As in the one-link model above, fares are differentiated if the inverse demand
functions of \( D_l \) and \( D_s \) differ, or if \( \beta \neq 1 \); in this model, this means that differentiation takes
place on at least one of the two links.

To measure whether the composition of users on each link is optimal, given a total number
of users, we again define a \( \Delta A \), as the change in welfare when the number of users from one
group is increased, while the total number of users is kept constant. There are now two of
these total usage levels, one on each link; both are kept constant if
\[ \Delta A = \frac{\partial W}{\partial N_l} - \frac{\partial W}{\partial N_s} \]

Using the monopolistic fares and first-order conditions, \( \Delta A = 2N_s D_s' - N_l D_l' \). Hence, in
general, \( \Delta A \neq 0 \); a more restrictive policy can increase social welfare, even if does not change
the usage levels on both links, or even decreases them. Of course, as usual in a monopoly,
usage levels are also lower than optimal; increasing the total number of users on each link
will also increase welfare.

2.3.3 Monopoly without price differentiation

If the monopolist cannot discriminate between users traveling between different OD-pairs, it
has to charge each user traveling on the same link the same price \( f_s \), and hence, \( f_l = 2f_s \).
This situation is sometimes called ‘link-based pricing’; see chapters 3, 5, and 6 for examples.
The monopolist then maximizes \( \pi = 2f_s (N_s + N_l) \) s.t.

\[ D_s (N_s) - c(\beta N_l + N_s) - f_s = 0 \]  \hspace{1cm} (2.34)
\[ D_l (N_l) - 2\alpha c(\beta N_l + N_s) - 2f_s = 0 \]  \hspace{1cm} (2.35)

The profit-maximizing fare is given by
\[ f = (N_l + N_s) \ell' \frac{D_l' + 2\alpha \beta D_s'}{D_l' + 2D_s' - 2\ell'(\alpha - 1)(\beta - 1)} \]
\[ -N_l D_l' \frac{D_s'}{D_l' + 2D_s' - 2\ell'(\alpha - 1)(\beta - 1)} \]
\[ -N_s D_s' \frac{D_l'}{D_l' + 2D_s' - 2\ell'(\alpha - 1)(\beta - 1)} \]
which is almost the same expression as for our first model; the only difference being the double
weight on \( D_l' \) in the denominator of all three elements (as one long-distance traveler could
Price differentiation and discrimination in transport networks

replace two short-distance travelers), and in the numerator of the first element. Naturally, this will generally cause fares to be discriminatory on at least one of the links.

Under uniform pricing, \( \Delta A = 2 (\alpha N_l + N_s) (1 - \beta) c' \). This means that, if \( \beta < 1 \), there too many short-distance travelers on both links, while if \( \beta > 1 \), there are too many long-distance travelers. Uniform pricing may have a positive or negative allocation effect, depending on the model parameters. If \( \beta = 1 \), the allocation of users is optimal; in that case, uniform pricing always has a positive allocation effect, as without uniform pricing the allocation of user classes is generally suboptimal from a societal point of view.

Determining what happens to the total number of users on each link is somewhat more complicated than for the one-link model above. However, it is still possible to define a fare difference \( \phi = 2f_2 - f_1 \), and analyze the impact of an increase in \( \phi \) on the usage levels \( N_s + N_l \). The derivations can be found in the appendix but, perhaps surprisingly, the results are exactly the same as those obtained for the one-link model above: in a linear case

\[
\frac{d (N_s + N_l)}{d\phi} = \frac{(\alpha - \beta) c'}{2\Omega} \tag{2.37}
\]

where, here,

\[
\Omega = \begin{vmatrix}
D'_s - c' & -\beta c' \\
-2\alpha c' & D'_l - 2\alpha\beta c'
\end{vmatrix} > 0 \tag{2.38}
\]

and hence, the total change in usage levels as a result of uniform pricing is given by \( (2f_2 - f_1) (\alpha - \beta) c' / 2\Omega \). In a nonlinear case, the expressions are slightly different than in a linear case, as the cross-derivatives of the usage levels with respect to both fares are assigned different weights, but they are qualitatively the same. In the same way, it is possible to show that, even if the two short-distance inverse demand functions are different, and if the two links have different cost functions, \( \phi \) still has no impact on the usage level of either of the two links as long as \( \alpha = \beta \) (see Appendix).

This has several implications for the welfare effects of uniform pricing. Firstly, and most importantly, the network context has no qualitative impact here: if all users’ values of time are equal to their congestion coefficients, uniform pricing does not change the usage levels of the individual links. In other words, OD-based price discrimination in a network is exactly the same as price differentiation between user classes traveling on the same link. This may sound surprising, but it is important to remember that it is not the same as saying that the total number of users \( in the network \) remains constant. The latter is obviously not the case if there is also an allocation effect.

This also means that, as before, output can be higher under uniform pricing than when differentiation is allowed. Here, this happens if \( (2f_2 - f_1) (\alpha - \beta) < 0 \), or, substituting in the unrestricted monopolistic fares, if \( \alpha - \beta \) and \( 2 (1 - \beta) (\alpha N_l + N_s) c' - 2N_s D'_s + N_l D'_l \) have opposite signs. There is nothing in the model that prevents this from happening; it is solely determined by the model parameters.

2.3.4 Monopoly without price discrimination

If the monopolistic operator is not allowed to discriminate, it maximizes the same profit function as in an unrestricted setting, subject to the same user equilibrium constraints, as well as a non-discrimination condition:

\[
2f_s - f_l + 2 (\beta - 1) (\alpha N_l + N_s) c' = 0 \tag{2.39}
\]
If this equality does not hold, price discrimination occurs on at least one of the links. The resulting fares are equal to

\[ f_s = (\alpha N_1 + N_s) c' - N_s D'_s \frac{D'_l - (3\alpha - 2\beta - \alpha\beta) c'}{D'_l + 2D'_s + c' (4 + \alpha) (\beta - 1)} \]

\[ f_l = 2(\alpha N_1 + N_s) \beta c' - 2N_s D'_s \frac{D'_l - (3\alpha - 2\beta - \alpha\beta) c'}{D'_l + 2D'_s + c' (4 + \alpha) (\beta - 1)} \]

assuming, for tractability, that cost functions are linear such that \( c'' = 0 \). Although the expression for the uniform fare in this model was very similar to its corresponding expression in the first model above, the difference is larger here; in the weights multiplying the two markups, the cost function is multiplied by a much more complex function of \( \alpha \) and \( \beta \). The structure of the expressions is still the same: they consist of each class’ marginal costs, plus a weighted average monopolistic markup. If \( \beta \leq 1 \), such that long-distance travelers have a lower relative congestion coefficient, the average markup is closer to the unrestricted long-distance markup if \( \alpha \) is lower; this is also likely, though not guaranteed, to happen if \( \beta > 1 \). As before, these non-discriminatory fares always lead to differentiation on at least one of the two links if \( \beta \neq 1 \).

As in the one-link model, the non-discrimination condition ensures that \( \Delta A = 0 \). The output effect is ambiguous in sign, and not equal to zero even in a linear world where \( \alpha = \beta \), as the fare difference \( \phi \) is endogenous.

Because this setting is so very similar to the previous, we will not present a numerical example. Instead, we directly turn to the third setting.

### 2.4 Network effects - parallel links

#### 2.4.1 Social optimum

In this model, there is only one inverse demand function, \( D(N_1 + N_2) \). However, users can now take two routes, where each route has its own average user cost; \( c_1(N_1) \) and \( c_2(N_2) \), respectively. Note that this is similar to some of the two-period models available in the literature (see e.g., Liu and McDonald, 1999); here, the two alternatives are perfect substitutes. Social welfare is then given by

\[ W = \int_0^{N_1+N_2} D(n) \, dn - N_1 c_1(N_1) - N_2 c_2(N_2) \]  

(2.42)

A social planner chooses \( f_1 \) (for route 1) and \( f_2 \) (for route 2) to maximize \( W \), subject to two user equilibrium constraints that ensure that sum of the average user costs and the fare is the same for both routes, and equal to the marginal user benefits:

\[ D(N_1 + N_2) - c_1(N_1) - f_1 = 0 \]  

(2.43)
2 Price differentiation and discrimination in transport networks

\[ D (N_1 + N_2) - c_2(N_2) - f_2 = 0 \] (2.44)

and hence, naturally, \( f_1 = N_1 c'_1 \) and \( f_2 = N_2 c'_2 \)

### 2.4.2 Unrestricted monopoly

A monopolist maximizes \( f_1 N_1 + f_2 N_2 \) s.t. the same two constraints as the social planner. Hence,

\[ f_1 = N_1 c'_1 - (N_1 + N_2) D' \] (2.45)

\[ f_2 = N_2 c'_2 - (N_1 + N_2) D' \] (2.46)

These fares are always non-discriminatory: \( f_1 - N_1 c'_1 = f_2 - N_2 c'_2 \); a non-discriminatory policy would not change anything. Moreover, the distribution of users over the two links is always optimal, as \( \Delta A \equiv \partial W/\partial N_1 - \partial W/\partial N_2 = 0 \). This means that a more restrictive policy always has a negative allocation effect; uniform pricing can only increase welfare if it increases the total number of users.

### 2.4.3 Monopolistic pricing without differentiation

In this setting, \( f_1 = f_2 = f \). Hence, the operator maximizes \( f(N_1 + N_2) \) s.t. \( D (N_1 + N_2) - c_1(N_1) - f = 0 \) and \( D (N_1 + N_2) - c_2(N_1) - f = 0 \)

This gives

\[ f = N_1 c'_1 - c_1 + c_2 + N_2 c'_2 - c_1 + c_2 - (N_1 + N_2) D' \] (2.47)

In contrast to the two models above, the operator still charges the same demand-related markup on both routes; this is natural, as there is only one demand function. The cost functions are different, so, instead of charging users of each route their marginal cost, the operator charges a weighted average marginal cost, with weights determined by the relative marginal costs. This time, the weights do sum to one: users only differ in one dimension, so only the part of the fare that is related to marginal costs needs to be averaged across groups. There are no further distortions, and hence, the monopolistic markup does not need any correction. If \( N_1 c'_1 \neq N_2 c'_2 \) (which will happen if \( c_1(\cdot) \neq c_2(\cdot) \)), these fares are discriminatory.

Again, we consider the effects of a non-differentiated pricing policy by separately looking at the allocation and output effects. In this model, \( \Delta A = -N_1 c'_1 + N_2 c'_2 \), so, in general \( \Delta A \neq 0 \): the distribution of users over the two links is not optimal. As we have established, an unrestricted monopoly does lead to an optimal composition, so uniform pricing always has a negative allocation effect.

As before, we can determine the effect of price differentiation on the total number of users by defining a fare difference \( \phi = f_2 - f_1 \). Totally differentiating the direct demand function \( N (f_1, f_1 + \phi) \equiv N_1 (f_1, f_1 + \phi) + N_2 (f_1, f_1 + \phi) \) gives

\[ \frac{dN}{d\phi} = \frac{\partial N}{\partial f_1} \frac{df_1}{d\phi} + \frac{\partial N}{\partial f_2} \frac{df_2}{d\phi} = \left( \frac{\partial N}{\partial f_1} + \frac{\partial N}{\partial f_2} \right) \frac{df_1}{d\phi} + \frac{\partial N}{\partial f_2} \] (2.48)
2.4 Network effects - parallel links

Totally differentiating the user equilibrium conditions and using Cramer’s rule gives expression for $\partial N/\partial f_1$ and $\partial N/\partial f_2$; substituting these in the expression above gives

$$\frac{dN}{d\phi} = -\frac{1}{\Omega} \left( (\epsilon'_1 + \epsilon'_2) \frac{\partial f_1}{\partial \phi} + \epsilon'_1 \right)$$

where

$$\Omega \equiv \left| \begin{array}{cc} D' - \epsilon'_1 D & \epsilon'_1 D' \\ D & -\epsilon'_2 \end{array} \right| > 0$$

(2.49)

(2.50)

This means that, if $\frac{\partial f_1}{\partial \phi} > 0$, the total number of users decreases in the amount of differentiation (and hence, that a uniform pricing policy increases output). If $\frac{\partial f_1}{\partial \phi} < -1$, a uniform pricing policy decreases output. More generally, as before

$$\frac{\partial f_1}{\partial \phi} = -\frac{\partial^2 \pi / (\partial \phi \partial f_1)}{\partial^2 \pi / (\partial f_1)^2}$$

(2.51)

where the nominator can be written as

$$\frac{\partial^2 \pi}{(\partial \phi \partial f_1)} = -2\epsilon'_1 + f_1 \left( \frac{\partial N}{\partial f_1 \partial f_2} + \frac{\partial^2 N}{(\partial f_2)^2} \right) + \phi \left( \frac{\partial (\partial N_1)^2}{\partial f_1 \partial f_2} + \frac{\partial^2 N_1}{\partial^2 f_2} \right)$$

(2.52)

and the denominator is positive. In a linear case,

$$\frac{\partial f_1}{\partial \phi} = -\frac{1}{2} \frac{\partial^2 \pi}{\partial \phi \partial f_1} = -\frac{\partial^2 f_1}{\partial \phi^2}$$

(2.53)

and hence, $dN/d\phi = 0$; there is no output effect. As the allocation effect is negative, uniform pricing always decreases welfare. If the demand function is non-linear, the second and third terms in Eq. 2.52 are nonzero, and the output effect may be positive or negative, depending on the Hessian of the direct demand functions.

2.4.4 Monopolistic pricing without discrimination

As already mentioned before, non-discrimination constraints would not be binding here; the unrestricted monopoly is already non-discriminatory. Discrimination only occurs if different users are charged a different monopolistic markup, which, in turn, can only happen if they have different demand structures.

2.4.5 Numerical example

Fig. 2.5 shows the results of a numerical simulation in which $c_1 = 10 + 0.1N_1$, $c_2 = 1.5 + C \cdot N_2$ and $D = 100 - (N_1 + N_2)$. As before, the vertical axis shows the relative efficiency of, in this case, the uniform pricing policy, compared to an unrestricted monopoly. Naturally, $\omega$ is always negative: in this linear example, $N_1 + N_2$ is the same, regardless of whether prices are differentiated or uniform, while the socially optimal division of travelers over the two links is optimal in the unrestricted monopoly. As $C$ increases, the number of users on the second link becomes smaller, and the uniform fare $f$ is set closer to what $f_1$ would be in an unrestricted monopoly. This reduces the allocative inefficiency of the uniform pricing policy; eventually,
2 Price differentiation and discrimination in transport networks

if $C$ becomes large enough, no travelers would take the second link, and price differentiation would not affect welfare at all.

2.5 Conclusions

In this chapter, we have analyzed price discrimination and price differentiation in transport networks. As we have shown, it is important to make the distinction between differentiation and discrimination in this context, especially if users have different values of time, or marginal costs differ for other reasons.

Our models confirm that enforcing uniform or non-discriminatory pricing policies can, in some circumstances, improve social welfare, but decrease welfare in others. Importantly, although a non-discriminatory policy may be considered the most ‘fair’ by users, it may be worse for welfare than uniform pricing, even if users have different marginal costs.

Generalizing the existing literature, we have examined a situation in which users not only have different marginal private costs (e.g., as a result of different values of time), but also different marginal external costs (e.g., because they are driving different passenger-car equivalents). This does matter: in a linear world, for instance, price differentiation still affects total usage levels even if all users have the same value of time. Only if each user’s relative value of time is equal to its relative marginal external cost does this effect disappear. Non-discrimination, on the other hand, always improves the composition of users, but may increase or decrease the total number even in a linear model where each user’s relative value of time is equal to its relative marginal cost, as marginal external costs depend on usage levels.

In addition to this analysis of price discrimination and discrimination on a single link, we have also considered situations in which there are parallel or serial links, and transport operators can differentiate and/or discriminate based on the route users take, or on their origin. In a network with serial links, a monopolistic operator may be able to discriminate based on users’ origins and destinations. As we have shown, this type of discrimination is not qualitatively different than discrimination based on values of time or marginal external costs. Although the monopolist’s fare setting rules are different, all conclusions obtained in a one-link model generalize to a network with serial links.

Route-based discrimination, as may occur in networks with parallel links, is different than OD-based discrimination. If demand is linear, enforcing uniform pricing over multiple routes can never increase welfare. More generally, uniform pricing can only increase welfare if it
substantially increases the usage level, because, if the only difference between users if the route they take, an unrestricted monopolist will set fares such that the division of users over the routes is optimal. An unregulated monopoly is already non-discriminatory if chosen route is the only difference between users.

Naturally, all three models we have examined are highly stylized. Any real-world network will simultaneously have parallel and serial links, and many different user classes traveling along them, between different OD-pairs. However, our stylized models do highlight the need for careful, situation-based analysis to evaluate the potential benefits of restrictive pricing policies. Neither uniform nor non-discriminatory pricing policies are universally welfare-enhancing (as first-degree price discrimination usually is) or universally decrease welfare (as a single-link linear model where all users have the same value of time would suggest). They also illustrate the need to distinguish between differentiation and discrimination. Although non-discrimination may be perceived as more ‘fair’, it is not always the best.

Our results also have implications for network modeling. Because of its computational advantages, link-based pricing, where operators do not differentiate based on OD-pairs, but charge all users the same for use of a link, is often assumed. If real-world transport operators are able to charge OD-based fares, these link-based models may understate or overstate the benefits of other types of regulation.

2.6 Appendix

2.6.1 Output effects in a model with two serial links

Symmetric model

Define \( \phi = 2f_s - f_l \), and direct, link-based demand \( N(f_s, 2f_s - \phi) \equiv N_s + N_l \). Then,

\[
\frac{dN}{d\phi} = \frac{\partial N}{\partial f_s} \frac{df_s}{d\phi} + \frac{\partial N}{\partial f_l} \frac{df_l}{d\phi}
\]

\[
= \frac{df_s}{d\phi} \left( \frac{\partial N}{\partial f_s} + 2 \frac{\partial N}{\partial f_l} \right) - \frac{\partial N}{\partial f_l}
\]

Totally differentiating the two user equilibrium conditions and applying Cramer’s Rule gives

\[
\frac{\partial N_s}{\partial f_s} = \frac{(D'_s - 2\alpha\beta\epsilon')}{\Omega}
\]
\[
\frac{\partial N_l}{\partial f_s} = \frac{2\alpha\epsilon'}{\Omega}
\]
\[
\frac{\partial N_s}{\partial f_l} = \frac{\beta\epsilon'}{\Omega}
\]
\[
\frac{\partial N_l}{\partial f_l} = \frac{(D'_l - \epsilon')}{\Omega}
\]

where \( \Omega = \begin{vmatrix} D'_s - \epsilon' & -\beta\epsilon' \\ -2\alpha\epsilon' & D'_l - 2\alpha\beta\epsilon' \end{vmatrix} > 0 \)
2 Price differentiation and discrimination in transport networks

Hence,
\[
\frac{\partial N}{\partial f_s} = \left(D'_l - 2\alpha (\beta - 1) c'\right) / \Omega
\]
\[
\frac{\partial N}{\partial f_l} = \left(D'_s + (\beta - 1) c'\right) / \Omega
\]

Finally,
\[
\pi (f_s) = 2f_s N_s (f_s, 2f_s - \phi) + (2f_s - \phi) N_l (f_s, 2f_s - \phi)
\]
\[
\pi' = 2N + 2f_s \left( \frac{\partial N}{\partial f_s} + \frac{\partial N}{\partial f_l} \right) - \phi \left( \frac{\partial N_l}{\partial f_s} + 2 \frac{\partial N_l}{\partial f_l} \right)
\]

Treating \( \phi \) as exogenous and totally differentiating this first-order condition with respect to \( \phi \) gives
\[
\frac{\partial f_s}{\partial \phi} = -\frac{\partial^2 \pi / (\partial \phi \partial f_s)}{\partial \pi / (\partial f_s)^2}
\]

where
\[
\frac{\partial^2 \pi}{(\partial \phi \partial f_s)} = 2 \frac{\partial N}{\partial f_s} \frac{\partial N_i}{\partial f_l} - 2 \frac{\partial N_i}{\partial f_l} + 2f_s \left( \frac{\partial^2 N}{(\partial \phi \partial f_s)^2} + 2 \frac{\partial^2 N}{(\partial \phi \partial f_l)^2} \right) - \phi \left( \frac{\partial^2 N_l}{(\partial f_s)^2} + 2 \frac{\partial^2 N_l}{\partial f_s \partial f_l} \right)
\]
\[
= -2 \frac{\partial N}{\partial f_l} - 2 \frac{\partial N_i}{\partial f_l} + 2f_s \left( \frac{\partial^2 N_l}{(\partial f_s)^2} + 2 \frac{\partial^2 N_l}{\partial f_s \partial f_l} \right) + \phi \left( \frac{\partial^2 N_l}{(\partial f_s)^2} + 2 \frac{\partial^2 N_l}{\partial f_s \partial f_l} \right)
\]

and
\[
\frac{\partial^2 \pi / (\partial f_s)^2}{(\partial f_s)} = \frac{4 \frac{\partial N}{\partial f_s} + 8 \frac{\partial N_i}{\partial f_l} + 2f_s \left( \frac{\partial^2 N}{(\partial f_s)^2} + 2 \frac{\partial^2 N_l}{\partial f_s \partial f_l} \right)}{\phi} - \phi \left( \frac{\partial^2 N_l}{(\partial f_s)^2} + 2 \frac{\partial^2 N_l}{\partial f_s \partial f_l} \right)
\]

If demand- and cost functions are linear,
\[
\frac{\partial f_s}{\partial \phi} = \frac{2D'_s + (\alpha + \beta - 2) c'}{2 \left(D'_l + D'_s - (2\alpha - 1) (\beta - 1) c'\right)}
\]
\[
\frac{dN}{d\phi} = \frac{1}{\Omega} \left[ \frac{(2D'_s + (\alpha + \beta - 2) c') (D'_l + 2D'_s - 2(\alpha - 1) (\beta - 1) c') - (D'_s + (\beta - 1) c')}{2 \left(D'_l + D'_s - (2\alpha - 1) (\beta - 1) c'\right)} \right]
\]
\[
\frac{dN}{d\phi} = \frac{(\alpha - \beta) c'}{2\Omega}
\]

General model

In a more general model, with three inverse demand functions \( D_{12} (N_{12}) \), \( D_{23} (N_{23}) \), \( D_{13} (N_{13}) \), and cost functions \( c_{12} \) and \( c_{23} \) for the short-distance travelers, and \( \alpha (c_{12} + c_{23}) \) for the long-distance travelers, it can be shown, in exactly the same way, that the output effect of an increase in \( \phi \) is still zero if \( \alpha = \beta \).

In this case, there are three fares, and the price difference \( \phi \) can be written as
\[
\phi \equiv f_{12} + f_{23} - f_{13} \iff f_{13} = f_{12} + f_{23} - \phi
\]
and there link-based total direct demand functions are given by

\[ N_1 (f_{12}, f_{23}, f_{12} + f_{23} - \phi) \equiv N_{12} + N_{13} \]

\[ N_2 (f_{12}, f_{23}, f_{12} + f_{23} - \phi) \equiv N_{23} + N_{13} \]

The impact of price differentiation on the usage level of first link can then be written as

\[
\frac{dN_1}{d\phi} = \frac{\partial N_1}{\partial f_{12}} \frac{df_{12}}{d\phi} + \frac{\partial N_1}{\partial f_{23}} \frac{df_{23}}{d\phi} + \frac{\partial N_1}{\partial f_{13}} \frac{df_{13}}{d\phi} \\
= \frac{\partial N_1}{\partial f_{12}} \frac{df_{12}}{d\phi} + \frac{\partial N_1}{\partial f_{23}} \frac{df_{23}}{d\phi} + \frac{\partial N_1}{\partial f_{13}} \left( \frac{df_{12}}{d\phi} + \frac{df_{23}}{d\phi} - 1 \right) \\
= \left( \frac{\partial N_1}{\partial f_{12}} + \frac{\partial N_1}{\partial f_{13}} \right) \frac{df_{12}}{d\phi} + \left( \frac{\partial N_1}{\partial f_{23}} + \frac{\partial N_1}{\partial f_{13}} \right) \frac{df_{23}}{d\phi} - \frac{\partial N_1}{\partial f_{13}}
\]

In this case, there are three user equilibrium conditions. Totally differentiating those, and using Cramer’s Rule gives:

\[
\frac{\partial N_{12}}{\partial f_{12}} = \left( (D'_{23} - c'_{23}) (D'_{13} - \alpha\beta c'_{12} - \alpha\beta c'_{23}) - \alpha\beta c'_{12} \right) / \Omega < 0 \\
\frac{\partial N_{23}}{\partial f_{12}} = \alpha\beta c'_{12} / \Omega < 0 \\
\frac{\partial N_{13}}{\partial f_{12}} = \left( D'_{23} - c'_{23} \right) (\alpha c'_{12}) / \Omega > 0 \\
\frac{\partial N_{12}}{\partial f_{23}} = \alpha\beta c'_{12} c'_{23} / \Omega < 0 \\
\frac{\partial N_{23}}{\partial f_{23}} = \left( (D'_{12} - c'_{12}) (D'_{13} - \alpha\beta c'_{12} - \alpha\beta c'_{23}) - \alpha\beta c'_{12} \right) / \Omega < 0 \\
\frac{\partial N_{13}}{\partial f_{23}} = \left( D'_{12} - c'_{12} \right) (\alpha c'_{23}) / \Omega > 0 \\
\frac{\partial N_{12}}{\partial f_{13}} = \left( D'_{23} - c'_{23} \right) (\beta c'_{12}) / \Omega > 0 \\
\frac{\partial N_{23}}{\partial f_{13}} = \left( D'_{12} - c'_{12} \right) (\beta c'_{23}) / \Omega > 0 \\
\frac{\partial N_{13}}{\partial f_{13}} = \left( D'_{12} - c'_{12} \right) (D'_{23} - c'_{23}) / \Omega < 0
\]

where \( \Omega \equiv \left| \begin{array}{ccc} D'_{12} - c'_{12} & 0 & -\beta c'_{12} \\
0 & D'_{23} - c'_{23} & -\beta c'_{23} \\
-\alpha c'_{12} & -\alpha c'_{23} & D'_{13} - \alpha\beta (c'_{12} + c'_{23}) \end{array} \right| < 0 \)

Finally, the first-order conditions for profit maximization can be written as
Price differentiation and discrimination in transport networks

\[
\begin{align*}
\frac{\partial \pi}{\partial f_{12}} &= N_1 + f_{12} \left( \frac{\partial N_1}{\partial f_{12}} + \frac{\partial N_1}{\partial f_{13}} \right) + f_{23} \left( \frac{\partial N_2}{\partial f_{12}} + \frac{\partial N_2}{\partial f_{13}} \right) - \phi \left( \frac{\partial N_{13}}{\partial f_{12}} + \frac{\partial N_{13}}{\partial f_{13}} \right) = 0 \\
\frac{\partial \pi}{\partial f_{23}} &= N_2 + f_{12} \left( \frac{\partial N_1}{\partial f_{23}} + \frac{\partial N_1}{\partial f_{12}} \right) + f_{23} \left( \frac{\partial N_2}{\partial f_{23}} + \frac{\partial N_2}{\partial f_{13}} \right) - \phi \left( \frac{\partial N_{13}}{\partial f_{23}} + \frac{\partial N_{13}}{\partial f_{13}} \right) = 0
\end{align*}
\]

treating \( \phi \) as exogenous and totally differentiating these first-order conditions gives

\[
\begin{align*}
\frac{d}{d \phi} \frac{\partial \pi}{\partial f_{12}} &= \frac{\partial^2 \pi}{\partial f_{12} \partial f_{12}} \frac{df_{12}}{d \phi} + \frac{\partial^2 \pi}{\partial f_{23} \partial f_{12}} \frac{df_{23}}{d \phi} + \frac{\partial^2 \pi}{\partial \phi \partial f_{12}} = 0 \\
\frac{d}{d \phi} \frac{\partial \pi}{\partial f_{23}} &= \frac{\partial^2 \pi}{\partial f_{12} \partial f_{23}} \frac{df_{12}}{d \phi} + \frac{\partial^2 \pi}{\partial f_{23} \partial f_{23}} \frac{df_{23}}{d \phi} + \frac{\partial^2 \pi}{\partial \phi \partial f_{23}} = 0
\end{align*}
\]

Using Cramer's Rule:

\[
\frac{df_{12}}{d \phi} = \frac{|H|}{\partial^2 \pi (\frac{df_{12}}{d \phi})^2} \quad \frac{df_{23}}{d \phi} = \frac{|H|}{\partial^2 \pi (\frac{df_{23}}{d \phi})^2}
\]

where \( H \) is the Hessian of \( \pi \):

\[
|H| = \begin{vmatrix}
\frac{\partial^2 \pi}{\partial f_{12}^2} & \frac{\partial^2 \pi}{\partial f_{12} \partial f_{23}} \\
\frac{\partial^2 \pi}{\partial f_{12} \partial f_{23}} & \frac{\partial^2 \pi}{\partial f_{23}^2}
\end{vmatrix} = \frac{\partial^2 \pi}{\partial f_{12}^2} \frac{\partial^2 \pi}{\partial f_{23}^2} - \left( \frac{\partial^2 \pi}{\partial f_{12} \partial f_{23}} \right)^2
\]

If demand- and cost functions are linear,

\[
\begin{align*}
\frac{\partial \pi}{\partial f_{12}} &= -\left( \frac{\partial N_1}{\partial f_{12}} + \frac{\partial N_{13}}{\partial f_{12}} + \frac{\partial N_{13}}{\partial f_{13}} \right) \\
\frac{\partial \pi}{\partial f_{23}} &= -\left( \frac{\partial N_1}{\partial f_{23}} + \frac{\partial N_{13}}{\partial f_{23}} + \frac{\partial N_{13}}{\partial f_{13}} \right) \\
\frac{\partial^2 \pi}{\partial f_{12} \partial f_{12}} &= 2 \left( \frac{\partial N_1}{\partial f_{12}} + \frac{\partial N_1}{\partial f_{13}} \right) \\
\frac{\partial^2 \pi}{\partial f_{23} \partial f_{12}} &= \frac{\partial N_1}{\partial f_{23}} + \frac{\partial N_1}{\partial f_{13}} + \frac{\partial N_2}{\partial f_{12}} + \frac{\partial N_2}{\partial f_{13}} \\
\frac{\partial^2 \pi}{\partial f_{23} \partial f_{23}} &= 2 \left( \frac{\partial N_2}{\partial f_{23}} + \frac{\partial N_2}{\partial f_{13}} \right)
\end{align*}
\]

These expressions can be substituted in \( \partial f_{12}/\partial \phi \) and \( \partial f_{23}/\partial \phi \), and the result combined with the partial derivatives of the usage levels with respect to the fares and the expression for \( dN_1/d\phi \) given above. This straightforward substitution exercise is too tedious even for this appendix, and the resulting expression is too long to print here. It is, however, equal to zero when \( \alpha = \beta \).
3 Competition in multi-modal transport networks: a dynamic approach\textsuperscript{1}

3.1 Introduction

The previous chapter dealt with the behavior of a monopolistic transport operator, and specifically focused on situations in which that operator charged different groups of users different prices. Although some transport markets are indeed monopolistic, others operate under varying levels of competition. This chapter therefore considers a competitive setting, in which rail operators compete with each other and with parallel roads, and compares this to a monopoly.

A substantial literature exists on road pricing, as well as on the effects of road pricing on public transport markets, and second-best pricing of roads in the presence of unpriced substitutes. In reality, road pricing is politically difficult to implement and, with a few exceptions, roads remain unpriced. It is therefore interesting to consider the opposite, but common situation, in which roads are unpriced, but public transport has a usage fee, which is set while taking the effects on the roads into account.

Recent decades have seen a shift from governmental provision of public transport to provision by private firms, thus reducing the government’s control over fares. However, not all systems have been privatized in the same way and consequently, various market structures have emerged. Although, in many countries, there is now at least some form of vertical separation between service operators and infrastructure managers, the amount of competition between service operators differs greatly. In some countries, such as The Netherlands, one operator owns the exclusive rights to operate most or even all connections in the network; in others, such as the UK, new operators can freely enter the market to offer new services, or directly compete for franchises to operate existing ones.

We therefore test how these different market structures influence public transport fares and social welfare, and how they compare with governmental provision. In particular, we analyze the difference between markets with a monopolistic public transport operator, which operates all public transport links, and markets in which separate operators own each public transport link. Standard economic theory would predict that in normal markets, both Bertrand (price) and Cournot (quantity) competition lead to lower prices than a monopoly; in the former, prices would be driven down to marginal costs when there are at least two competitors; in the latter, prices approach marginal costs only when the number of competitors approaches infinity.

Transport systems usually consist of several interacting markets, and there are unpriced externalities associated with travel, so these standard results do not always apply. Previous studies (Economides and Salop, 1992; see De Borger et al., 2008 for a transport application)

\textsuperscript{1}This chapter is based on joint work with Erik Verhoef and Vincent van den Berg. An earlier version of this chapter has been published in *Transportation Research Part B* (van der Weijde et al., 2013b). I thank Hugo Silva, Sergej Gubins and the participants of the 2011 Kuhmo Nectar conference for their helpful comments and suggestions.
have shown that, generally, parallel (or horizontal) competition, where a number of competitors offer different possibilities to travel between two points, is beneficial, both to consumers and to society. In contrast, serial (or vertical) competition, in which different operators own complementary links, increases fares and adversely affect social welfare. In this situation, each serial competitor exerts local market power, and is able to set a price above marginal cost. If the overall demand is price-sensitive, this has a negative effect on the patronage of other links, but this externality is disregarded by the individual operators; a phenomenon that is comparable to the mechanism of double marginalization.

However, these results have been obtained with static models. In these models, commuters only choose a mode, or combination of modes, to travel; the models disregard the fact that commuters can also choose the moment at which they travel. In most real-world applications, commuters do have this choice, and empirical evidence suggests that this has non-negligible effects (Small, 1982). In contrast to the previous chapter, we will therefore use dynamic modeling techniques, and examine how this affects competition. In order to do so, we assume that demand is fixed, such that we can isolate dynamic interaction effects from possible effects of price-sensitive demand.

To further improve tractability, we only examine networks in which commuters from different origins, located along one transportation corridor, travel to one destination; this may, for example, represent a morning commute from a series of suburbs to a central business district. More general network models exist (e.g. Pels and Verhoef, 2007), but are often too complicated to yield the economic insights we are interested in. Simpler multi-modal network models (e.g. Arnott and Yan, 2000; Verhoef, 2008) often assume that there are only two nodes in the network, that only one mode is chosen for the entire journey, or that all parallel links are exactly the same, which is too simple for our purposes. There is also some earlier literature on the properties of congested many-to-one commuter networks similar to ours (e.g. Tian et al., 2007; Arnott and DePalma, 2011), but these are usually concerned with the user equilibrium only, and do not include competing parallel modes. Most recently, Li et al. (2012) analyzed a multi-modal many-to-one transport corridor in which modes share the same highway, but their model does not allow for serial competition between public transport operators.

We therefore combine elements from different contributions to this literature, to model a simple dynamic multi-modal many-to-one commuter network where transfers between modes are possible, but costless. In this way, we can capture the essence of serial competition, parallel competition, mode choice and departure time choice in one analytical framework, and study the efficiency of different types of market organization. Using this model, we obtain a reduced form of the public transport operator’s optimal fare setting problem, assuming that its fare is constant over time. This reduced form is considerably simpler and easier to use than the original optimal control problem, and we can use it to show that in this dynamic model, even though the total travel demand is inelastic, serial Bertrand-Nash competition on the public transport links leads to different fares than a serial monopoly. However, contrary to the results obtained in classic studies on vertical competition, the difference between duopolistic and monopolistic fares is not necessarily positive. We also show why these results cannot be observed in static models, and further examine these results in a series of numerical simulations.

The following section will outline the methodology and assumptions. In section 3.3, we examine the fare-setting behavior of public transport operators in a static model with fixed demand, and briefly discuss why the monopolistic and serial Nash-Bertrand equilibrium fares
3.2 Methodology

Even our simple models require a significant amount of notation. Table 3.1 summarizes the main indexes, variables and parameters that will be used in our exposition. To examine the effects of different market structures in a multi-modal network, we will first consider the simplest possible network in which this is possible, and later consider how the results obtained can be generalized. This simple network, shown in Fig. 3.1, consists of three nodes (two origins and one destination), which are connected by two segments. Each segment consists of two links. We will call these two links “rail” and “road”, but they could also represent other modes, as long as they are completely separate, such that commuters using one node do not influence the travel costs of the commuters on the other node. Importantly, we ignore the discrete nature of public transport, and instead assume that passengers can depart and arrive continuously. Although it is theoretically possible to include a limited number of departures, this major complication would not qualitatively affect the results. Both modes are congestible; roads, because travel speeds depend on the number of passengers, and trains, because travelers experience discomfort as a result of in-vehicle crowding (see Wardman and Whelan, 2011 and Li and Hensher, 2011 for an overview of the literature on crowding in public transport).

Travelers treat the two modes as perfect substitutes, so that Wardropian equilibrium conditions apply for used alternatives; their generalized prices should be equal in equilibrium, and there are no unused alternatives with a lower price (Wardrop, 1952). If modes were imperfect substitutes, this would reduce the effects of competition, but it would not eliminate them.

We examine a typical morning commute, in which \( N \) commuters travel from node 1 to the destination, and another \( N \) commuters from node 2 to the destination. Commuters from node 1 can transfer to a different mode at node 2, although, as we will see in sections 3 and 4 below, this assumption does not influence the results in an interior equilibrium; it is also possible to disallow all transfers or only allow transfers in one direction. Transfers are costless since, in an interior equilibrium, a positive transfer cost would simply eliminate all transfers. The total number of commuters traveling from each node, \( N \), is fixed, as we want to exclude the effects of elastic demand from our analysis, and in particular keep them from complicating the comparison of equilibrium use levels across alternative market configurations.

As a benchmark, we first consider a static model, in which departure time decisions are ignored; the congestion costs commuters face are influenced by all commuters using the same links. Afterwards, we will examine a dynamic model, in which commuters choose a departure time to minimize the total cost of traveling, and only the number commuters traveling on

---

2 Especially if passengers do not know the schedule and headways are exogenous; in that case, each passenger would simply incur an additional cost, equal to the value of waiting time multiplied by the expected waiting time. If passengers know the schedule or headways are endogenous, solutions will be very difficult to obtain (See also Tian et al., 2007).

3 This assumed equality of commuter numbers simplifies notation, but could otherwise easily be dropped.
### Indices

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>Segments (1, 2)</td>
</tr>
<tr>
<td>m</td>
<td>Modes (R = road, T = train)</td>
</tr>
<tr>
<td>t</td>
<td>Time</td>
</tr>
</tbody>
</table>

### Variables

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c_{\text{mon}} )</td>
<td>Operation and maintenance costs of a monopolistic operator</td>
</tr>
<tr>
<td>( c_l )</td>
<td>Operation and maintenance costs of a duopolistic operator on segment ( l )</td>
</tr>
<tr>
<td>( f_{l}^{m} )</td>
<td>Usage fee for mode ( m ) on segment ( l ) (road price or rail fare)</td>
</tr>
<tr>
<td>( n_{l}^{m} )</td>
<td>Number of commuters traveling on mode ( m ), segment ( l )</td>
</tr>
<tr>
<td>( t_{1}^{m} )</td>
<td>Arrival time at the destination of the first commuter from the first node who has used mode ( m )</td>
</tr>
<tr>
<td>( t_{2}^{m} )</td>
<td>Arrival time at the destination of the first commuter from the second node who has used mode ( m )</td>
</tr>
<tr>
<td>( v_{l}^{m} )</td>
<td>Arrival flow on link ( l ), mode ( m )</td>
</tr>
<tr>
<td>( r_{l}^{m} )</td>
<td>Congestion costs faced by users of mode ( m ) on segment ( l )</td>
</tr>
<tr>
<td>( s_{l}^{m} )</td>
<td>Travel speed on link ( l ), mode ( m )</td>
</tr>
<tr>
<td>( \pi_{\text{mon}} )</td>
<td>Profit of a monopolistic operator</td>
</tr>
<tr>
<td>( \pi_l )</td>
<td>Profit of a duopolistic operator on segment ( l )</td>
</tr>
<tr>
<td>( \theta )</td>
<td>Schedule delay costs of a commuter arriving at the destination</td>
</tr>
</tbody>
</table>

### Parameters

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>Number of commuters traveling from each node to the destination</td>
</tr>
<tr>
<td>( M )</td>
<td>Segment length</td>
</tr>
<tr>
<td>( t^* )</td>
<td>Desired arrival time (common to all travelers)</td>
</tr>
</tbody>
</table>

Table 3.1: Notation
the same links at the same time influences the congestion cost they face. In that model, there is flow congestion on the road, as in Chu (1995), a reformulation of Henderson (1974); travel speeds of commuters depend on the traffic flow at the moment of arrival, where there are no specific restrictions on the functional form of this relation. This assumption allows us to obtain reduced-form formulations that could not have been obtained if other forms of congestion, such as bottleneck congestion, would have been present.

In the general formulation, we only examine interior solutions, in which all modes are used from all origins; where necessary, we assume that such a solution exists. The numerical analysis allows for all possible equilibria, and thus for the possibility that one or more links may remain unused in equilibrium.

3.3 Static model

3.3.1 Generalized prices

Commuters traveling over segment $l$ pay a road price $f_R^l$ or rail fare $f_T^l$, where the index $T$ stands for train travel, and $R$ for road travel. We will assume that these fares are time-invariant, and that operators cannot discriminate between users traveling on a given link. In addition to that, they also face a congestion cost $r_{ml}^m$, $m \in \{T, R\}$, which is a function of the number of commuters traveling on that same link, $n_{ml}^m$. Since there is no time dimension in the static model, schedule delay costs are not defined.

3.3.2 Rail monopoly

A monopolistic rail operator, which operates both rail segments, maximizes the sum of the revenues of both links minus operating and maintenance costs $c_{mon}$:

$$
Max \quad \pi_{mon} = f_T^1 n_T^1 + f_T^2 n_T^2 - c_{mon} \left( n_T^1, n_T^2 \right)
$$

(3.1)

There are two Wardropian user equilibrium constraints for commuters from the upstream node, as they can follow three possible mode choice patterns (train–train, car–car and car–train), and one constraint ensuring user equilibrium for commuters from the downstream node, as they can only choose between two modes (train or car). Two other constraints apply, ensuring that all commuters from each origin travel. The Lagrangian for profit opti-
Combining and simplifying the first-order conditions yields the optimal monopolistic rail operation and maintenance costs may be larger than those of the monopolist if the latter can

where \( \lambda_1 \) to \( \lambda_4 \) are the Lagrange multipliers associated with the constraints mentioned above. Combining and simplifying the first-order conditions yields the optimal monopolistic rail prices:

\[
f_l^T = \frac{\partial c_{mon}}{\partial n_l^T} + n_l^T \frac{\partial r_l^T}{\partial n_l^T} + n_l^T \frac{\partial r_l^R}{\partial n_l^R}
\] (3.3)

Hence, the price on each link consists of the marginal cost of accommodating an extra commuter on that link, and a congestion-related mark-up. This mark-up consists of two parts. The second term in Eq. 3.3 is the increase in congestion costs that all rail users experience as a result of an increase in the number of users; naturally, the monopolist internalizes these costs. The third term in Eq. 3.3, which also internalizes part of the congestion on the road, is the equivalent of a standard monopolistic mark-up in a market with price-sensitive demand. With an inverse demand function \( d_l^R(n_l^T) \) and in absence of an alternative mode, this mark-up would be equal to –\( n_l^T \left( \frac{\partial d_l^R}{\partial n_l^T} \right) \). Because, in our model, any rail traveler shifted away from rail to road, the term \( \partial r_l^R \partial n_l^R \) in Eq. 3.3 is equivalent to what \( -\partial d_l^R / \partial n_l^T \) would be in this conventional setting.

### 3.3.3 Rail Bertrand-Nash duopoly

If the two rail links are owned by separate operators, each operator maximizes its own profits, while both maximization problems are subject to the same constraints as those applying for the monopolistic operator. Hence, the two Lagrangians are:

\[
L_{b1} = f_1^T n_1^T - c_1 \left( n_1^T \right) + \lambda_{b1} \left[ r_1^T - r_1^R \left( n_1^R \right) - f_1^R \right]
\] (3.4)

\[
L_{b2} = f_2^T n_2^T - c_2 \left( n_2^T \right) + \gamma_{b1} \left[ r_1^T - r_1^R \left( n_1^R \right) - f_1^R \right]
\] (3.5)

where \( \lambda_{b1} \) to \( \lambda_{b4} \) and \( \gamma_{b1} \) to \( \gamma_{b4} \) are Lagrangian multipliers. The sum of the two duopolists’ operation and maintenance costs may be larger than those of the monopolist if the latter can
exploit economies of scale, so \( c_1 + c_2 \geq c_{mon} \). The first-order conditions yield:

\[
\frac{f_T}{c_{mon}} = \frac{\partial c_1}{\partial n_T} + n_T \frac{\partial c_T}{\partial n_T} + n_R \frac{\partial c_R}{\partial n_T}
\]  

(3.6)

With the exception of the first term, Eq. 3.6 is equal to Eq. 3.3; contrary to a standard Bertrand duopoly, both operators can raise their prices above marginal production cost \( c_T^r \), for two reasons. The first is that they internalize the congestion externality on their own link, and collect the revenue from the associated toll. The second is that the serial setup implies that they are not offering pure substitutes, but rather complements, so that the parallel road is their direct competition; not the other duopolist.

### 3.3.4 Price comparison

If the monopolistic operator does not benefit from economies of scale in the number of rail links it owns, and thus \( \frac{\partial c_1}{\partial n_T} = \frac{\partial c_{mon}}{\partial n_T} \), Eqs. 3.3 and 3.6 are equal; there is no difference between Bertrand-Nash and monopolistic prices. Only if \( \frac{\partial c_1}{\partial n_T} > \frac{\partial c_{mon}}{\partial n_T} \), which may be the case if a monopolistic operator can save costs as a result of owning both links, it is possible that monopolistic and duopolistic fares are different. Since the exact parameters of the model determine whether this happens or not, we will ignore this case in the analyses below.

To some extent, the above results are intuitive. Since demand is perfectly inelastic, all users have to travel, and all commuters from node 1 (the upstream node) pass through node 2, where they can take the train regardless of the mode they used to arrive there. Hence, the number of train travelers between each pair of nodes does not depend on the rail fare on the other link. In other words, a duopolistic operator does not have to take into account how its customers arrived at the start of its segment or how they will depart at the end, nor how much this costs the customers, since the number of customers arriving at and departing from each segment is fixed. The monopolistic optimization problem is therefore perfectly separable in the two segments, and the two duopolists set exactly the same prices as the monopolist.

So far, we have allowed commuters from node 1 to transfer between modes in either way, by not placing any restrictions on \( n_T^2 \) and \( n_R^2 \). However, transfers between modes do not have to be considered explicitly; it can be shown mathematically that the inclusion of an additional constraint specifying that \( n_T^2 > n_T^1 \) (such that only transfers from road to train are possible) or \( n_T^2 > n_T^1 \) and \( n_R^2 > n_R^1 \) (such that no transfers are possible) does not change the results, as these constraints do not affect the rail fares, and will apply to both duopolists and to the monopolist. The reason that the fare setting problem is still separable, even when transfers are not possible, is that commuters from the downstream node equalize the generalized price of travel on both modes on the downstream segment. Hence, when commuters from the upstream node decide which mode to use, it is sufficient for them to only consider the road prices and rail fares on the upstream segment, since the full price of traveling on the downstream segment is, thanks to the downstream commuters, independent of the chosen mode.

### 3.4 Dynamic model

The static model above does not allow commuters to choose a departure time that minimizes their generalized travel price. Although this may be realistic in some settings, departure
time choices cannot be ignored in most real-world situations, especially in the context of
congestion, and it is important to consider if the model outcomes are different if departure
time choices are indeed included.

The full dynamic formulation of the problem outlined above yields a complex optimal con-
trol problem, as rail operators should now maximize the time integral of profits on each link,
subject to integral constraints and temporal Wardropian equilibrium conditions. However, it
is possible to derive a reduced-form formulation, under the assumption that fares are constant
over time, as is the case in many real-world situations. In order to do this, we first establish
the arrival order of commuters from the different origins and modes. We then divide the
time between the arrival of the first and last commuters at the destination in four periods,
and subsequently solve the resulting reduced-form model as if it were a static optimization
problem.

3.4.1 Generalized prices and arrival order

As before, users face road prices or rail fares, and a congestion cost $r_m^l$. In the dynamic
model, however, the congestion costs are a function of the number of commuters arriving at
the same time, so $r_m^l(t) = r_m^l(v_m^l(t))$, where $v_m^l(t)$ is the number of commuters using mode
$m$ arriving at time $t$ at the end of link $l$. In addition, commuters also face a schedule delay
cost $\theta$, which increases as commuters arrive further away from their preferred arrival time
$t^*$. For simplicity, we assume that late arrivals (after $t^*$) are not allowed.\footnote{4} We also assume
that the free-flow travel speed of a car is higher than the speed of a train; this allows for
establishment of the arrival order.\footnote{5}

Define $t^R_1$ as the time the first road–road commuter from node 1 arrives at the destination,
and $t^R_2$ the time the first road commuter from node 2 arrives. To ensure user equilibrium,
$\frac{\partial r^R_1}{\partial v^R_1} \frac{\partial v^R_1}{\partial t^R_1} + \frac{\partial r^R_1}{\partial v^R_1} \frac{\partial \theta}{\partial t} = 0$ when commuters from node 1 are traveling, and
$\frac{\partial r^R_2}{\partial v^R_2} \frac{\partial v^R_2}{\partial t^R_2} + \frac{\partial \theta}{\partial t} = 0$ when commuters from node 2 are traveling, such that generalized prices are constant over
time and no commuters want to change their arrival time. Moreover, generalized prices faced
by users from both nodes have to be lower when they are traveling then when they are not
traveling.

Using these facts, it can be shown that $t^R_2 > t^R_1$, that the flow of road-road commuters from
node 1 is positive from $t^R_1$ to $t^*$, and that the flow of road commuters from node 2 is positive
from $t^R_2$ to $t^*$. Moreover, as soon as road commuters from node 2 start to arrive, the arrival
flow of commuters from node 1 becomes constant.

These properties are not difficult to prove. Commuters from node 2 increase their arrival
flow over time to exactly offset the decrease in their schedule delay costs. If commuters from
node 1 travel at the same time, the flow on the first segment must be constant to keep the
generalized prices for these commuters constant over time. This also implies that commuters
from node 1 must start to travel first, such that their arrival flow is strictly positive when
commuters from node 2 start to travel.\footnote{6}

\footnote{4}It is also possible to allow late arrivals, which is more realistic, but this complicates the analysis without
yielding additional insights. In line with some of the previous literature, (e.g. Arnott and Kraus, 1995,
1993; Kraus and Yoshida, 2002) late arrivals are therefore prohibited.

\footnote{5}In the absence of a road price, this assumption could be relaxed; the first road user would depart earlier
than the first rail user even if the free-flow speeds of both these modes were equal.

\footnote{6}Tian et al. (2007) more rigorously show some of these properties for a single-mode many-to-one network
with discrete departure times. Arnott and DePalma (2011) do the same for a single-mode many-to-one
network where origins are distributed continuously.
3.4 Dynamic model

<table>
<thead>
<tr>
<th></th>
<th>Road RI</th>
<th>$t_1^R \leq t &lt; t_2^R$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Road RII</td>
<td>$t_2^R \leq t \leq t^*$</td>
</tr>
<tr>
<td></td>
<td>Train TI</td>
<td>$t_1^T \leq t &lt; t_2^T$</td>
</tr>
<tr>
<td></td>
<td>Train TII</td>
<td>$t_2^T \leq t \leq t^*$</td>
</tr>
</tbody>
</table>

Table 3.2: Time periods

Define $t_1^R$ as the time the first rail–rail commuter from node 1 arrives at the destination, and $t_2^R$ the time the first rail commuter from node 2 arrives. Assuming that the free-flow travel speed on the road is higher than or equal to the travel speed of the train, $t_1^R \geq t_1^T$. As on the road, the first upstream commuter arrives first, so $t_1^R > t_1^T$. Hence, $t_1^R \leq t_1^T$, $t_2^R$, $t_2^T < t^*$. Fig. 3.2 shows an example of the user equilibrium arrival flows in model with linear congestion costs and linear costs of schedule delay.

Using these insights, we can divide period between the first and last arrival into four partially overlapping periods; two for road travel and two for rail travel, as summarized in Table 3.2. Within each time period, the generalized price faced by commuters using the indicated mode is constant, to satisfy the Wardropian user equilibrium constraints.

Figure 3.2: Equilibrium arrival order and arrival flows at destination
3 Competition in multi-modal transport networks

3.4.2 Price setting – monopolistic rail operator

Period RI

As explained above, the generalized user price for road-road commuters from the upstream node arriving at the destination at time $t$ is given by:

$$f_1^R + f_2^R + r_1^R \left( v_1^R \left( t - \frac{M}{s_2^R (t)} \right) \right) + r_2^R \left( v_2^R (t) \right) + \theta (t; \beta) \quad (3.7)$$

Note that commuters arriving at the destination at time $t$ enter the second segment at time $t - \frac{M}{s_2^R (t)}$, as $\frac{M}{s_2^R (t)}$ is the time needed to traverse the second segment; hence, congestion costs, which are assumed to be a function of the flow of commuters out of the segment, are incurred at that time.

In a Wardropian equilibrium, these generalized prices are time-invariant, so the time-derivative of Eq. 3.7 is equal to zero during the whole period. Moreover, since no commuters from another node use the road in this period and both links are identical, each individual commuter’s speed is constant along the whole route, so $v_1^R \left( t - \frac{M}{s_2^R (t)} \right) = v_2^R (t)$, and therefore

$$\frac{\partial v_1^R \left( t - \frac{M}{s_2^R (t)} \right)}{\partial t} = \frac{\partial v_2^R (t)}{\partial t} = - \frac{\partial \theta (t)}{\partial t} \left( \frac{\partial r_1^R}{\partial v_1^R} + \frac{\partial r_2^R}{\partial v_2^R} \right) \quad (3.8)$$

An expression for the road travel flow in this period can be obtained by integrating Eq. 3.8:

$$v_1^R \left( t - \frac{M}{s_2^R (t)} \right) = v_2^R (t) = - \int \frac{\partial \theta (t)}{\partial t} \left( \frac{\partial r_1^R}{\partial v_1^R} + \frac{\partial r_2^R}{\partial v_2^R} \right) dt \quad (3.9)$$

where, by construction, $v_1^R \left( t_1^R - \frac{M}{s_2^R (t_1^R)} \right) = v_2^R (t_1^R) = 0$; there a no arrivals before $t_1^R$. Using this, the total numbers of commuters using both roads during this interval, $N_1^{RI}$ (for the first segment) and $N_2^{RI}$ (for the second segment) can be written as the integral of the arrival flow, which is now a function of $t_1^R$ and $t_2^R$ only:

$$N_1^{RI} = N_2^{RI} = \int_{t_1^R}^{t_2^R} v_2^R (t) = N_1^{RI} \left( t_1^R, t_2^R \right) dt \quad (3.10)$$

Period TI

In this period, the generalized user costs for train–train commuters are constant over time. Moreover, since only commuters from the upstream node use the train, $v_1^T \left( t - \frac{M}{s_1^T (t)} \right) = v_2^T (t)$. $v_1^T$ and $v_2^T$ are therefore given by:

$$v_1^T \left( t - \frac{M}{s_1^T (t)} \right) = v_2^T (t) = - \int \frac{\partial \theta (t)}{\partial t} \left( \frac{\partial r_1^T}{\partial v_1^T} + \frac{\partial r_2^T}{\partial v_2^T} \right) dt \quad (3.11)$$
where, by construction, \( v_1^T \left( t_1^T - \frac{M}{s_2^T (t_1^T)} \right) = v_2^T \left( t_1^T \right) = 0 \). Hence, the total number of commuters traveling by train on the first and second segment, respectively, can be written as:

\[
N_1^{TII} = N_2^{TII} = \int_{t_1^T}^{t_2^T} v_2^T (t) = N_1^{TII} \left( t_1^T, t_2^T \right) \, dt \tag{3.12}
\]

**Period RII**

Now, the generalized user costs for road–road commuters from the upstream node and road commuters from the downstream node are constant over time, and equal to those of users in the previous period. Moreover, the flow of road–road commuters from the upstream node is constant, so \( \frac{\partial v_R^1 \left( t - \frac{M}{s_2^R (t)} \right)}{\partial t} = 0 \). Hence, the total number of commuters using the first road segment in this period can be obtained through multiplication of the commuter flow at \( t_R^2 - \frac{M}{s_2^R (t_R^2)} \) (since commuters exiting the first segment at that time will arrive at their destination at \( t_R^2 \)) with the length of the period:

\[
N_1^{RII} = \left( \left( t^* - \frac{M}{s_2^R (t^*)} \right) - \left( t_R^2 - \frac{M}{s_2^R (t_R^2)} \right) \right) v_1^R \left( t_R^2 - \frac{M}{s_2^R (t_R^2)} \right) \tag{3.13}
\]

On the downstream link, \( \frac{\partial v_R^2 (t)}{\partial t} = -\frac{\partial \theta (t)}{\partial t} \frac{\partial v_R^2}{\partial v_R^2} \), and therefore

\[
v_2^R (t) = -\int \frac{\partial \theta (t)}{\partial t} \frac{\partial v_R^2}{\partial v_R^2} \, dt \tag{3.14}
\]

where \( v_2^R (t_R^2) \) is given by Eq. 3.9. Thus,

\[
N_2^{RII} = \int_{t_R^2}^{t^*} v_2^R (t) \, dt = N_2^{RII} \left( t_1^R, t_2^R \right) \tag{3.15}
\]

**Period TII**

Here, the generalized user costs for all types of commuters are constant over time, and equal to those of users in the previous period. Moreover, \( \frac{\partial v_T^1 \left( t - \frac{M}{s_2^T (t)} \right)}{\partial t} = 0 \). Hence,

\[
N_1^{TII} = \left( \left( t^* - \frac{M}{s_2^T (t^*)} \right) - \left( t_T^2 - \frac{M}{s_2^T (t_T^2)} \right) \right) v_1^T \left( t_T^2 - \frac{M}{s_2^T (t_T^2)} \right) \tag{3.16}
\]

\[
= N_1^{TII} \left( t_1^T, t_2^T \right)
\]
where \( \bar{v}_2^T \) is given by Eq. 3.11.

### 3.4.3 Monopolistic rail fares

A monopolistic operator chooses the four earliest arrival times \( \{t^R_1, t^R_2, t^T_1, t^T_2\} \) and both fares to maximize the sum of profits in all four periods:\(^7\)

\[
\max \pi_m = (N^T_1 + N^{TI}_2) f^T_1 + (N^T_1 + N^{TI}_2) f^T_2
\]

Since we have expressed the numbers of commuters using each mode as a function of the arrival times of all groups of commuters, using the Wardropian user equilibrium conditions, these four arrival times are now decision variables, instead of the original commuter numbers. Since fares are time-invariant, no further temporal user equilibrium constraints are necessary, because they have already been substituted in the total commuter numbers. However, two intermodal user equilibrium constraints are needed to ensure that, for users from both nodes, the generalized prices of both modes are equal. These constraints can be evaluated at a number of points in time, but, since all groups of commuters are traveling at \( t = t^* \), it is convenient to use that point:

\[
f^T_1 + r^T_1 (\bar{v}^T_1 (t^*_1, t^*_2)) = f^R_1 + r^R_1 (\bar{v}^R_1 (t^*_1, t^*_2))
\]

\[
f^T_2 + r^T_2 (\bar{v}^T_2 (t^*_1, t^*_2)) = f^R_2 + r^R_2 (\bar{v}^R_2 (t^*_1, t^*_2))
\]

where \( \bar{v}^m_l \) is the flow on mode \( m \), link \( l \) at time \( t = t^* \). The final two constraints ensure that all commuters travel:

\[
N^T_1 + N^{TI}_1 + N^R_1 + N^{RI}_1 = N
\]

\[
N^T_2 + N^{TI}_2 + N^R_2 + N^{RI}_2 = 2N
\]

The Lagrangian for profit maximization then becomes:

\[
L_m = \pi_m + \lambda_1 \left( f^T_1 + r^T_1 (\bar{v}^T_1) - f^R_1 - r^R_1 (\bar{v}^R_1) \right)
\]

\[
+ \lambda_2 \left( f^T_2 + r^T_2 (\bar{v}^T_2) - f^R_2 - r^R_2 (\bar{v}^R_2) \right)
\]

\[
+ \lambda_3 \left( N^T_1 + N^{TI}_1 + N^R_1 + N^{RI}_1 - N \right)
\]

\[
+ \lambda_4 \left( N^T_2 + N^{TI}_2 + N^R_2 + N^{RI}_2 - 2N \right)
\]

where \( \lambda_1 \) to \( \lambda_4 \) are the Lagrangian multipliers associated with constraints 3.19–3.22. The relevant first-order conditions are:

\[
\frac{\partial L_m}{\partial f^T_1} = N^T_1 + N^{TI}_1 + \lambda_1 = 0
\]

\(^7\)For simplicity, we assume that costs have no influence on the difference between the optimal monopolistic and duopolistic prices, and thus that there are no economies of scale in the number of links owned. If this were not the case, general conclusions are impossible (see also the static model above).
3.4 Dynamic model

\[ \frac{\partial L_m}{\partial f^T_2} = N_2^{TI} + N_2^{TII} + \lambda_2 = 0 \]  
(3.25)

\[ \frac{\partial L_m}{\partial t^R_1} = -\lambda_1 \frac{\partial r^T_1}{\partial v^R_1} \frac{\partial v^R_1}{\partial t^R_1} - \lambda_2 \frac{\partial r^T_2}{\partial v^R_2} \frac{\partial v^R_2}{\partial t^R_1} + \lambda_3 \left( \frac{\partial N_1^{RI}}{\partial t^R_1} + \frac{\partial N_1^{RII}}{\partial t^R_1} \right) \]  
(3.26)

\[ + \lambda_4 \left( \frac{\partial N_2^{RI}}{\partial t^R_1} + \frac{\partial N_2^{RII}}{\partial t^R_1} \right) = 0 \]

\[ \frac{\partial L_m}{\partial t^T_1} = \frac{f^T_1 \left( \frac{\partial N_1^{TI}}{\partial t^T_1} + \frac{\partial N_1^{TII}}{\partial t^T_1} \right) + f^T_2 \left( \frac{\partial N_2^{TI}}{\partial t^T_2} + \frac{\partial N_2^{TII}}{\partial t^T_2} \right)}{\partial t^T_1} \]  
(3.27)

\[ + \lambda_1 \frac{\partial r^T_1}{\partial v^T_1} \frac{\partial v^T_1}{\partial t^T_1} + \lambda_2 \frac{\partial r^T_2}{\partial v^T_2} \frac{\partial v^T_2}{\partial t^T_1} + \lambda_3 \left( \frac{\partial N_1^{TI}}{\partial t^T_1} + \frac{\partial N_1^{TII}}{\partial t^T_1} \right) = 0 \]

\[ \frac{\partial L_m}{\partial t^R_2} = -\lambda_1 \frac{\partial r^T_1}{\partial v^R_2} \frac{\partial v^R_2}{\partial t^R_2} - \lambda_2 \frac{\partial r^T_2}{\partial v^R_2} \frac{\partial v^R_2}{\partial t^R_2} + \lambda_3 \left( \frac{\partial N_1^{RI}}{\partial t^R_2} + \frac{\partial N_1^{RII}}{\partial t^R_2} \right) \]  
(3.28)

\[ + \lambda_4 \left( \frac{\partial N_2^{RI}}{\partial t^R_2} + \frac{\partial N_2^{RII}}{\partial t^R_2} \right) = 0 \]

\[ \frac{\partial L_m}{\partial t^T_2} = \frac{f^T_1 \left( \frac{\partial N_1^{TI}}{\partial t^T_2} + \frac{\partial N_1^{TII}}{\partial t^T_2} \right) + f^T_2 \left( \frac{\partial N_2^{TI}}{\partial t^T_2} + \frac{\partial N_2^{TII}}{\partial t^T_2} \right)}{\partial t^T_2} \]  
(3.29)

\[ + \lambda_1 \frac{\partial r^T_1}{\partial v^T_2} \frac{\partial v^T_2}{\partial t^T_2} + \lambda_2 \frac{\partial r^T_2}{\partial v^T_2} \frac{\partial v^T_2}{\partial t^T_2} + \lambda_3 \left( \frac{\partial N_1^{TI}}{\partial t^T_2} + \frac{\partial N_1^{TII}}{\partial t^T_2} \right) = 0 \]

Using Eqs. 3.24–3.29, it is possible to obtain closed-form solutions for \( f^T_1 \) and \( f^T_2 \); however, the resulting expressions are very tedious, and economic interpretation is therefore hard to give.

3.4.4 Duopolistic rail fares

A Bertrand operator on the first segment maximizes its own profit, subject to the same constraints as the monopolistic operator. The relevant first-order conditions for the two duopolists are very similar to the ones for the monopolist; Eqs. 3.24–3.26 and 3.28 are unchanged, although the shadow prices are different. Eqs. 3.27 and 3.29 are replaced by:

\[ \frac{\partial L_{b1}}{\partial t^T_1} = \frac{f^T_1 \left( \frac{\partial N_1^{TI}}{\partial t^T_1} + \frac{\partial N_1^{TII}}{\partial t^T_1} \right) + \gamma_1 \frac{\partial r^T_1}{\partial v^T_1} \frac{\partial v^T_1}{\partial t^T_1} + \gamma_2 \frac{\partial r^T_2}{\partial v^T_2} \frac{\partial v^T_2}{\partial t^T_1} + \gamma_3 \left( \frac{\partial N_1^{TI}}{\partial t^T_1} + \frac{\partial N_1^{TII}}{\partial t^T_1} \right) + \gamma_4 \left( \frac{\partial N_2^{TI}}{\partial t^T_1} + \frac{\partial N_2^{TII}}{\partial t^T_1} \right)}{\partial t^T_1} = 0 \]  
(3.30)
3 Competition in multi-modal transport networks

\[
\frac{\partial L_{b1}}{\partial t_1} = f_1^T \left( \frac{\partial N_{11}^{TI}}{\partial t_1^1} + \frac{\partial N_{11}^{TI1}}{\partial t_1^1} \right) + \gamma_1 \frac{\partial n_1^1}{\partial v_1^1} \frac{\partial v_1^1}{\partial t_1^1} + \gamma_2 \frac{\partial n_1^2}{\partial v_2^2} \frac{\partial v_2^2}{\partial t_1^2} + \gamma_3 \left( \frac{\partial N_{11}^{TI}}{\partial t_2^1} + \frac{\partial N_{11}^{TI1}}{\partial t_2^1} \right) + \gamma_4 \left( \frac{\partial N_{21}^{TI1}}{\partial t_2^1} + \frac{\partial N_{21}^{TI1}}{\partial t_2^1} \right) = 0 \tag{3.31}
\]

\[
\frac{\partial L_{b2}}{\partial t_1} = f_2^T \left( \frac{\partial N_{21}^{TI}}{\partial t_1^1} + \frac{\partial N_{21}^{TI1}}{\partial t_1^1} \right) + \varphi_1 \frac{\partial n_1^1}{\partial v_1^1} \frac{\partial v_1^1}{\partial t_1^1} + \varphi_2 \frac{\partial n_1^2}{\partial v_2^2} \frac{\partial v_2^2}{\partial t_1^2} + \varphi_3 \left( \frac{\partial N_{11}^{TI}}{\partial t_1^2} + \frac{\partial N_{11}^{TI1}}{\partial t_1^2} \right) + \varphi_4 \left( \frac{\partial N_{21}^{TI1}}{\partial t_1^2} + \frac{\partial N_{21}^{TI1}}{\partial t_1^2} \right) = 0 \tag{3.32}
\]

\[
\frac{\partial L_{b1}}{\partial t_2} = f_2^T \left( \frac{\partial N_{21}^{TI}}{\partial t_2^2} + \frac{\partial N_{21}^{TI1}}{\partial t_2^2} \right) + \varphi_1 \frac{\partial n_1^1}{\partial v_1^1} \frac{\partial v_1^1}{\partial t_2^2} + \varphi_2 \frac{\partial n_1^2}{\partial v_2^2} \frac{\partial v_2^2}{\partial t_2^2} + \varphi_3 \left( \frac{\partial N_{11}^{TI}}{\partial t_2^2} + \frac{\partial N_{11}^{TI1}}{\partial t_2^2} \right) + \varphi_4 \left( \frac{\partial N_{21}^{TI1}}{\partial t_2^2} + \frac{\partial N_{21}^{TI1}}{\partial t_2^2} \right) = 0 \tag{3.33}
\]

Again, it is possible to solve for \( f_1^T \) and \( f_2^T \), but this does not yield helpful results.

3.4.5 Fare comparison and comparison with static model

Although, in this general setting, it is not possible to determine in which cases the fare differentials between the monopolistic and duopolistic settings is positive or negative, the difference between Eqs. 3.27–3.29 and 3.30–3.33 clearly shows that there is no reason that this differential will be zero, as it was in the static model. Each duopolist disregards the effect that a change in its fare will have on its competitor’s patronage. Fares on one link influence patronage on the other, because a change in one fare affects the trip timing of all commuters using that link, which indirectly affects the trip timing of all commuters, and therefore all generalized prices. A monopolist internalizes these effects.

This phenomenon is not unrelated to the ‘double marginalization’ in static networks with price-sensitive demand, but it is different. In that case, competitors reduce demand to a level below the social optimum, by charging fares that are too high. Here, competitors shift demand to sub-optimal times, by charging fares that, in principle, could be either too high or too low. Naturally then, a static model does not capture this effect, as it disregards trip timing, and hence only allows for inefficiencies in the total commuter flow rather than the flow at any point in time. If demand is fixed, the monopolist’s fare setting problem is no longer perfectly separable in two duopolistic problems if there are possibilities for intertemporal substitution. If, in addition, demand is price-sensitive, both the total number of commuters using each mode and their arrival times can be suboptimal; both the effect discussed above and the static ‘double marginalization’ are present.

There is, however, an important difference between the two. The static ‘double-marginalization’-effect always leads to duopolistic fares that are too high. If one of the two operators increases its price, that will always reduce its competitor’s profit; any fare increase carries a negative externality on the other operator’s profit. Here, the externality may also be positive, which would decrease duopolistic fares. To see this, consider the effect of an increase in \( f_1^T \) on the linear model in Fig. 3.2. Since this increase makes the train
more expensive, relative to the road, the first rail commuter will arrive later; $t_1^f$ will move to the right, decreasing the flow of rail–rail commuters from node 1 for all points in time. Rail commuters from the downstream node are not directly affected by this fare increase, but they are affected by the reduced flow of commuters from the upstream node. This reduces their costs, so their flow will increase, and they will start to arrive earlier; $t_2^f$ will move to the left. If the congestion and schedule delay functions are such that the change in $t_2^f$ is large, profits on the second segment may increase, rather than decrease.

An increase in $f_2^T$ has a different effect. Again, if the behavior of travelers from node 2 would not change, it makes rail travel more expensive, relative to the road, for commuters from node 1; they will arrive later, which reduces the incoming flow into node 2. However, rail commuters from that node now also face a higher fare, so $t_2^f$ may move to the left or right. If it moves far enough to the right, flows on the first segment may in fact continue to increase for longer than before, and hence, the first operator’s profits may increase. Again, this depends on the exact form of the congestion and schedule delay functions. It is more likely to happen if the schedule delay function has a steep slope, such that flows increase more rapidly, and a small change in $t_2^f$ has a larger effect on the total number of rail–rail commuters from node 1.

In Section 3.5, we analyze these effects in a numerical model. There, we do indeed find that the externality one operator imposes on its competitor can be both negative or positive, and hence, that duopolistic fares can be higher or lower than monopolistic fares, depending on the parameters of the model.

### 3.4.6 Towards a general network

The above results, and the reduced-form optimization problem used to obtain them, also apply to a more general multi-modal many-to-one commuter network. The only assumption needed is that, in equilibrium, all routes are used at least some time. As long as this is the case, adding more serial origin nodes, or adding more parallel modes between all nodes will not qualitatively change the results. It is also possible to allow the total number of commuters per node to differ across nodes. Although this will complicate the derivations, the qualitative results will again not change.

It is significantly more difficult to also allow for travel to more destinations; for example, from node 1 to 2. In this case, travel time growth on a downstream link cannot at the same time compensate for schedule delay cost developments for travelers using only that link, and travelers using that link together with other links, when travel delays are time-varying. This would lead to a temporal separation of travelers from different OD-pairs, and different models have to be developed to address this. It is also difficult to allow for other network structures, where not all nodes can be placed on a straight line; in that case, establishing an arrival order ex-ante may be impossible, and numerical simulations would have to be performed for every possible combination. Also for the present setting, however, numerical modeling can produce additional insights, as we will see in the next section.

### 3.5 Numerical analysis

The above analysis shows how a reduced-form of the dynamic user equilibrium can be derived analytically. This reduced form was then used to prove that, in a dynamic model, there is likely to be a difference between monopolistic and duopolistic fares. It is possible to obtain
3 Competition in multi-modal transport networks

closed-form expressions for the user equilibrium flows, as functions of the fares. However, these allow for little, if any, economic interpretation. Closed-form expressions for the optimal fares are more difficult to obtain, and even more difficult to interpret. It is not possible to derive when this difference is positive or negative, nor how exactly it is influenced by the various model parameters, without specifying the congestion functions $r^{ml}(v^{ml})$, and the schedule delay function $\theta(t)$. However, these questions are important, particularly to assess whether they lead to fares and commuter flows that are further away from the optimum, or whether they could cancel out some of the negative effects of the traditional static ‘double marginalization’-phenomenon in a more realistic network setting with price-sensitive demand.

We therefore present the results of a numerical analysis, and use those results to identify how the various model parameters influence optimal monopolistic and duopolistic fares. Although much of the previous literature has focused on fares only, it is equally important to examine the welfare implications of the two market structures. To do so, we also consider the relative social costs under both market structures; since demand is fixed, this is the negative of social surplus.

3.5.1 Functions and base case parameters

Train speeds, $S^T$, are constant. However, commuters face in-vehicle crowding costs, which we assume to increase linearly in the commuter flow. Fares are constant over time. A traveler’s total costs of traveling on a link are then:

$$r^T_l (t) + f^T_l = \alpha \frac{M}{S^T} + M \xi_1 v^T_l (t) + f^T_l$$  \hspace{1cm} (3.34)

where $\alpha$ is the value of travel time, $M$ the segment length, and $\xi_1$ a parameter. Link-based congestion costs for road users consist only of the costs of travel time and, as before, roads are not priced:

$$r^R_l (t) = \alpha \frac{M}{s^R_l (t)}$$  \hspace{1cm} (3.35)

where $s^R_l (t)$ is the road speed on segment $l$ at time $t$. Travel times increase linearly in the commuter flow\footnote{I.e., there is flow congestion as in Chu (1995). Since the network consists of two serial segments, the more popular bottleneck congestion is difficult to implement. For simplicity, the congestion function is linear; there is no hypercongestion.} with a maximum (free-flow) speed $S^R$:

$$\frac{1}{s^R_l (t)} = \frac{1}{S^R} + \xi_2 v^R_l (t)$$  \hspace{1cm} (3.36)

where $\xi_2$ is a parameter. Finally, schedule delay costs $\theta(t)$ are linear in $t$:

$$\theta(t) = \beta (t^* - t)$$  \hspace{1cm} (3.37)

where $\beta$ is the value of schedule delay early. As before, late arrivals are prohibited. Base case parameters are listed in Table 3.3; they are chosen to result in realistic fares and costs of congestion.

For simplicity, we assume that the marginal cost of transporting passengers is zero; assuming a constant positive marginal cost would complicate the analysis without qualitatively changing the results. Without loss of generality, fixed costs are also set to zero. The total
3.5 Numerical analysis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>5/hr</td>
</tr>
<tr>
<td>$\beta$</td>
<td>3/hr</td>
</tr>
<tr>
<td>$M$</td>
<td>10 km</td>
</tr>
<tr>
<td>$S^T$</td>
<td>100 km/h</td>
</tr>
<tr>
<td>$S^R$</td>
<td>120 km/h</td>
</tr>
<tr>
<td>$\xi_1$</td>
<td>0.002 (km/h)$^{-1}$</td>
</tr>
<tr>
<td>$\xi_2$</td>
<td>0.002 (pass./hr)$^{-1}$ (km/h)$^{-1}$</td>
</tr>
<tr>
<td>$N$</td>
<td>100 pass.</td>
</tr>
</tbody>
</table>

Table 3.3: Base case parameters

Social costs are then

$$C = \sum_{m} \sum_{l} \int r^m_l(t) + \theta(t) dt$$

$$= N \left( r^R_1(t^*) + 2r^R_2(t^*) \right) - \left( N^{TI}_1 + N^{TI}_2 \right) f^T_1 - \left( N^{TI}_2 + N^{TI}_2 \right) f^T_2$$

(3.38)

as generalized prices are equal on both modes, while only rail travelers pay a fare.

3.5.2 Value of time and value of schedule delay

We first vary the value of schedule delay, fixing all other parameters to the values listed in Table 3.3; in effect, we therefore vary the ratio between the value of schedule delay and the value of time. The left panel in Fig. 3.3 shows the difference between the duopolistic and monopolistic price for each segment, as a fraction of the monopolistic price. The right panel shows the social cost difference as a fraction of monopolistic welfare; positive values indicate that social costs are higher in a duopoly.

![Figure 3.3: Varying the value of schedule delay](image-url)
3 Competition in multi-modal transport networks

An increase in $\beta$ has two effects. Since being early at work is now more expensive, commuters will want to arrive closer to their desired arrival time, and the total network-wide congestion costs will be higher (see Eqs. 3.9 and further above, where now $\partial \theta(t)/\partial t = \beta$). Both the monopolist and the duopolists take advantage of the increased level of congestion to raise fares, but the duopolists disregard the effect on each other’s patronage, and consequently raise fares by a higher amount than the monopolist. This increases the fare differential. However, an increased value of schedule affects rail and road users in different ways. While, in this linear model, road travel flows increase at rate $\beta/2\alpha M \xi^2$ in period RI and twice as fast in period RII (see Eqs. 3.8 and 3.35–3.36), rail flows increase at rate $\beta/2M \xi^1$ in period TI and twice as fast in period TII (see Eqs. 3.11 and 3.34). Since rail flows increase at a higher rate to begin with, an increase in the value of schedule delay also has a higher impact on train users. Hence, an increase in $\beta$ reduces the market power of rail operators, and thus the tendency of the duopolists to charge higher fares, which reduces the price differential. As Fig. 3.3 shows, the second effect is clearly stronger on the more heavily congested downstream segment, as well as on the downstream segment for large values of $\beta$. The first effect dominates for very small values of $\beta$, but they do vary by large amounts.

The figure also shows that, whatever the value of schedule delay, the fare differential is much higher on the upstream segment. This results from the fact that all fares are lower on this segment due to the lower number of commuter traversing it; the same absolute fare difference therefore leads to much higher fractions on the upstream segment.

Finally, although the difference is small in relative terms, the duopolistic fare on the downstream segment can be lower than the monopolistic fare. This is more likely to happen as the value of schedule delay approaches the value of time, and implies that commuters from node 2 can be worse off in the monopolistic situation. This is an important observation, as it is a clear difference with the static ‘double marginalization’ effect; while the latter can only lead to higher duopolistic fares, the dynamic effect can reduce fares and hence, work in the opposite direction. However, although the numerical results therefore confirm the analytical finding that the duopolistic fare may be smaller than the monopolistic fare, this effect may have a limited quantitative impact in practice, despite its qualitative significance. The right panel in Fig. 3.3 also shows that, for society as a whole, a monopoly is still preferable. Although the relative social costs do vary with the value of schedule delay, as a result of varying fares and commuter flows, social costs remain between 4% and 5% higher in the duopolistic situation.

3.5.3 Congestion

Fig. 3.4 shows the results of a similar exercise as above, in which $\xi_1$, the rail congestion parameter, is varied from $1/1000$ to $1/100$. This parameter multiplies the commuter flow in the link-based rail cost functions, so an increase in $\xi_1$ increases the cost of in-vehicle crowding. Since $\xi_2$ is held constant, this increases the importance of congestion in the train relative to the importance of congestion on the road. We can therefore again expect two effects: firstly, as $\xi_1$ increases, rail operators face tougher competition from the road, which will decrease fares, and the tendency of duopolistic competitors to charge higher fares than a monopolist. However, the congestion costs increase in the whole network, as commuters equate the generalized prices of the two nodes. This increases the tendency of duopolistic competitors to charge higher fares.

All fares do indeed decrease in $\xi_1$; as congestion in the train is more costly, competition from
3.5 Numerical analysis

the road increases, and rail fares subsequently decrease. The relative fare difference between duopoly and monopoly, however, increases in $\xi_1$ on both segments; the total congestion effect is, in this case, stronger. Again, however, this effect is relatively small in the observed range. Again, also, a monopoly is always more efficient than a duopoly, although the cost difference between the two is very small for a very high unit cost of congestion. This shows that the social surplus effects of a parameter change can be rather different from the effects on fares.

3.5.4 Congestion and value of schedule delay

So far, we have considered the influence of the congestion parameter and the value of schedule delay separately. Of course, there are also interaction effects between these two parameters. Fig. 3.5 therefore shows the effects of a change in the value of schedule delay for a congestion parameter twice as high as the value used in Fig. 3.3 ($\xi_1 = 0.004$).

Comparing Figs. 3.3 and 3.5, it is clear that this interaction effect is important. With this higher congestion parameter, the first, direct effect of an increase in $\beta$ on congestion (see above) and thus on the ability of operators to charge higher fares, is much more important. However, the ranges of relative fare differences have not changed much. The influence of the value of schedule delay on the fare differential is still limited and social cost differences, although lower, also do not change by large amounts when $\beta$ is varied.

3.5.5 Commuter numbers

Fig. 3.6 shows the results of a simulation in which the number of commuters departing from each node was increased to up to ten times the base case value. Naturally, all fares increase in the number of commuters, as an increased number of commuters increases the congestion costs in the system. Both the monopolist and the duopolists are able to increase their fares as a result of a higher number of commuters. The duopolists, disregarding the externality they impose on their competitor’s patronage, increase their fares more than the monopolist; hence, absolute fare differences increase in $N$. However, the relative fare difference changes...
3 Competition in multi-modal transport networks

The relative fares and relative cost differences appear to be remarkably robust in this numerical exercise. That suggests that lessons on desirable market organization one can draw for this particular spatial setup are quite generic, which is good news if policies are to be developed in a changing world or under conditions of uncertainty on parameter values. Moreover, although duopolistic fares are smaller than monopolistic fares for at least some parameter ranges, this effect, though qualitatively significant, may only have a limited quantitative im-

3.5.6 Numerical conclusions

Figure 3.5: Varying the value of schedule delay ($\xi_1 = 004$)

Figure 3.6: Varying the number of commuters
3.6 Conclusions

We have shown that, in a multi-modal many-to-one commuter network where rail operators compete with unpriced roads, serial competition can influence fares, even in the absence of elastic demand. This results from the fact that, in a dynamic model, commuters have possibilities for intertemporal substitution, even if they do not have the option not to travel. Therefore, a fare change on one rail link changes not only the number of rail commuters relative to the number of road commuters, but also their trip timing decisions. If demand is fixed, the former will not necessarily affect anyone traveling on other links in the network, as we have shown with a simple static model, but the latter will change passengers flows in the whole network and through those, other operators’ patronages and profits. Naturally, a monopolistic operator internalizes the effect a price increase on one link has on the patronage of the other links, but a Bertrand-Nash operator disregards this, which leads to different, and potentially higher fares.

However, compared to static models, dynamic models are much more difficult to solve. For the many-to-one network described above, we have derived a reduced form under the assumption that fares are time-invariant, in which the rail operators optimize their fares and the boundaries of arrival time intervals, rather than the commuter flows. This makes subsequent calculations considerably easier, although general analytical solutions are still difficult to obtain.

Our numerical simulations show that, although duopolistic fares are higher than monopolistic for a wide range of parameters, this need not always be the case. Especially if the value of time approaches the value of schedule delay, and the unit cost of congestion in the train is higher than on the road, fares on at least one link may in fact be lower in a duopoly. This is a distinct difference with the static ‘double marginalization’-effect observed in models with price-sensitive demand. However, the simulations also show that model parameters, such as the value of schedule delay and the cost of congestion have only a limited effect on the fare difference between a monopoly and a duopoly and on the differences in social costs. This indicates that the conclusions drawn above are robust to assumption changes, and may also apply in many other spatial setups.

More research is necessary to explore these issues in at least three directions. Firstly, the effect of ownership on the substitute mode (the road) should be examined, especially in situations where the road is owned by a public operator, who sets a second-best road price to offset the negative effects of serial competition in the rail market. Secondly, as the numerical simulation shows that the fare increases may not affect all groups of commuters equally, local policy makers may try to influence the rail market to compensate the commuters in their area by, for example, subsidizing rail commuters from traveling from one node, or limiting the number of train services on one link. This could be modeled by adding an additional level, in which local governments maximize the social benefit in one node only by manipulating the variables over which they have control. Finally, a more general network setting, in which there is more than one destination, should be explored, as spatial interactions are clearly important.

However, despite the relative simplicity of the models outlined above, they do show that it...
3 Competition in multi-modal transport networks

is important to consider the effects of market structures in a situation where rail operators compete with an unpriced road, and to consider them in a dynamic setting. This introduces another reason for fare differences between different forms of competition, through the commuter’s departure timing choices. As we have shown, although these effects often work in the same direction as the classic ‘double marginalization’ that occurs when demand is elastic, there is a potential for it to work in the opposite direction. In any case, it should not be ignored, and we have shown ways to facilitate implementation.
4 A Hotelling model with price-sensitive demand and asymmetric distance costs: the case of strategic transport scheduling

4.1 Introduction

The Hotelling model is probably the most well-known model for studying product differentiation in markets with multiple competitors. Although originally framed in the context of locational choice along a linear market, it has various possible interpretations. One of these is to consider timing as the relevant measure of product differentiation, in which case Hotelling’s ‘space’ becomes ‘clock time’, and ‘transport costs’ between consumers and suppliers become ‘schedule delay costs’. In this chapter, we propose a general formulation of the Hotelling model that has exactly this interpretation.

In particular, we study how competing transport operators, active in the same market, can use the departure times of their services as strategic instruments. This has, so far, received relatively little attention in the literature; usually, only fares and frequencies are considered. In the previous chapters, scheduling decisions have also been ignored. However, a better understanding of strategic scheduling is important, not only because it could help explain observed changes in service stability as a result of deregulation (see, for example, Douglas, 1987), but also because departure time choices can have a large impact on transport users, and hence, regulators should know how to deal with them.

There is some existing literature on scheduling, mostly related to the deregulation of the British bus industry in the 1980s. These studies often assume an infinitely repeating schedule (for example, Foster and Golay, 1986; Evans, 1987), such that headways are constant, or focus on head-running and other short-term scheduling decisions (for example, Ellis and Silva, 1998). Most similar to our approach, van Reeven and Janssen (2006) analyze transport scheduling decisions using a circular Salop model. This model builds upon the Hotelling model and avoids its endpoints by making the market circular rather than having two end points that have the maximum distance in the market. In this way, schedules are still infinitely repeated. If there are no transport services during the day, or if there is a distinct off-peak period, this is less appropriate; moreover, van Reeven and Janssen need an additional attribute, such as service quality, for a stable equilibrium to emerge, and travelers have to care more about that attribute than about fares or departure times. This makes the model less attractive. We therefore propose a Hotelling model, in which operators schedule services in a discrete time interval.

Hotelling’s (1929) classic paper on horizontal differentiation argues that, when two firms compete on locations only, and a given number of consumers distributed along a linear market

---

1This chapter is based on joint work with Erik Verhoef and Vincent van den Berg. An earlier version of this chapter has been published in the Journal of Transport Economics and Policy (van der Weijde et al., 2014). I thank the participants of the 2012 Kuhmo Nectar conference and the 2014 ASSA Meetings for helpful comments.
buy from the closest firm, the two firms locate as closely together as possible. Later work has
generalized Hotelling’s model in several ways; most importantly (d’Aspremont et al., 1979)
have shown that that Hotelling’s original result was invalid if firms set prices, in addition to
locations. Price-setting behavior results in the absence of a pure-strategy equilibrium when
combined with linear transport costs, and to maximum differentiation when combined with
quadratic transport costs.

Most analyses have kept the assumptions that demand at every location is perfectly inelas-
tic, and that the user costs of traveling are independent of direction. There are exceptions;
Wauthy (1996b) formulates a two-stage model with elastic demand in the context of verti-
cal differentiation. Puu (2002) proposes a Hotelling model with elastic demand, in which
locations and prices are determined simultaneously (a mathematical formulation of Smithies,
1941), but his calculations have been shown to be flawed (Sanner, 2005). Colombo (2011)
includes elastic demand and asymmetric distance costs, but his model is unidirectional: travel
costs in one direction are infinite. Nilssen (1997) and Nilssen and Sørgard (2002) formulate lo-
cation choice models with asymmetric distance costs, but assume that prices are exogenously
fixed, and that there is only one consumer with unit demand at each location. Finally,
Gu and Wenzel (2012) formulate a Salop model with elastic demand, but their formulation
is unfortunate, in that the distance between consumers and their suppliers negatively af-
fects consumer utility, while their demand is a function of the price only; not of the transport
costs. To our knowledge, there are no existing horizontal differentiation models which include
price-sensitive demand at every location, allow for asymmetric distance costs, and include
price-setting in addition to location choice.

The assumptions mentioned above may yield good approximations in many applications,
but for transport scheduling, they are oversimplifications. Transport demand is usually price-
sensitive, since people can choose not to travel, or alternative modes of transport may be
available. Furthermore, Hotelling’s ‘distance’ or ‘travel costs’ in this setting represent the
costs of schedule delay, and the cost of being late is usually higher than the cost of being
early (Small, 1982). Our models generalize Hotelling’s horizontal differentiation model to in-
clude price-sensitive demand and allow for asymmetric schedule delay costs. Although we do
so in the context of transport scheduling, our models can be applied in other instances, such
as telecommunications markets (for example, as a generalization of Cancian et al., 1995).

As in a traditional horizontal differentiation model, two competitors choose a location on
a fixed interval. In our case, this is an interval in time, such that the two locations are
departure times of transport services. The two competitors also set their fares. Consumers
are distributed uniformly along the interval; their location indicates their desired departure
time, such that they face a schedule delay cost that increases in the deviation from their
desired departure time. Hence, they minimize their generalized price, which is the sum of the
fare and their schedule delay costs. Because we focus on longer-term scheduling decisions,
we assume that consumers know exactly when each service will depart, which avoids ‘bus
bunching’ and other short-term effects that rely on consumers taking being unable to use a
service that departs before their desired departure time.

Most Hotelling models assume that the two competitors choose their locations or departure
times first, after which fares are set. In scheduling, the opposite order is also conceivable,
but we show that this game does not have a Nash equilibrium; the same is true for a game in
which fares and departure times are chosen simultaneously. We find the equilibrium fares and
departure times, examine under which conditions equilibria exist, and also compare them to
the social optimum, assuming that schedule delay costs early and late are equal. Contrary
to other horizontal differentiation models, notably that of d’Aspremont et al. (1979), the competitors schedule their services closer together than optimal. We then analyze how these equilibria change if the schedule delay cost late is higher than the schedule delay cost early, and show that the resulting equilibria can still be stable. Finally, we analyze Stackelberg games, which increase the parameter space in which equilibria exist, and comment on optimal second-best regulation.

4.2 Methodology

Consider a departure time and fare choice game between two duopolistic suppliers of a scheduled transport service. Consumers have different preferences over desirable departure times and, given this desired departure time, there is an elastic demand for trips. Specifically, travel demand $d(t)$, at any time $t \in [-1, 1]$, is a linear function$^2$ of the generalized price $p(t)$. Since demand cannot be negative

$$d(t) = \text{Max} [a - bp(t), 0]$$

which holds for all $t$. The total consumer surplus can then be calculated as

$$CS = \frac{1}{2} \int_{-1}^{1} d(t) \left( \frac{a}{b} - p(t) \right) dt = \frac{1}{2b} \int_{-1}^{1} (d(t))^2 dt$$

Operators choose a fare $f_i$ and departure time $t_i$; this assumes that each operator schedules only one service in the observed time interval. Operator profits are given by

$$\pi_i = D_i f_i - F$$

where $D_i$ is the total demand for its service and $F$ is the fixed cost associated with the operation of the service; marginal per-passenger costs are assumed to be zero. Social welfare is then simply the sum of consumer surplus and operator profits, $CS + \sum_i \pi_i$.

The generalized price $p_i(t)$ of service $i$, taken by a consumer with preferred departure time $t$, is the sum of the fare and the schedule delay cost. Schedule delay costs are assumed to be linear, but not necessarily symmetric, such that the unit cost of schedule delay when a passenger is late ($\gamma$) can differ from the unit cost of schedule delay when a passenger is early ($\beta$):

$$p_i(t) = f_i + \text{Max} [\gamma (t_i - t), 0] + \text{Max} [\beta (t - t_i), 0]$$

Passengers choose the service that minimizes the generalized price they pay, so in equilibrium, their generalized price $p(t) = \text{Min}[p_1(t), p_2(t)]$. If the equilibrium demand for each operator’s service is strictly positive, we can define the inner market boundary point $t^*$, at which passengers are indifferent between the two operators as

$$t^* = (f_2 - f_1 + \beta t_1 + \gamma t_2)/(\beta + \gamma)$$

where, by construction, operator 1 schedules the first service. The total demand for each operator’s service, $D_1$ and $D_2$, can then be calculated as

$^2$In some Hotelling models, $t \in [0, 1]$, but the present specification yields more compact expressions, without affecting the conclusions.
4.3 Social optimum

Since there are no externalities associated with demand, and assuming that the operator’s marginal costs are zero, the social optimum can be found by simply maximizing total demand. This implies that both fares should be equal to zero. Assuming that, in the social optimum, \(d(-1), d(t^*), d(1) > 0\), meaning that there is positive demand from all preferred arrival times, maximizing the sum of Eqs. 4.6 and 4.7 gives the social welfare-maximizing departure times:

\[
\left\{ t_{SW1}^*, t_{SW2}^* \right\} = \left\{ -\frac{\gamma}{\beta + \gamma}, \frac{\beta}{\beta + \gamma} \right\}
\] (4.8)

This implies that, if \(\beta = \gamma\), the socially optimal departure times are at \(-1/2\) and \(1/2\). If the cost of departing late is higher than the cost of departing early, both departures shift to an earlier time, such that more passengers depart early. Note, however, that this assumes that demand is strictly positive for any \(t \in [-1, 1]\). Substituting the zero fares and Eq. 4.8 into Eq. 4.1, this implies that

\[
\frac{a}{b} - \frac{\beta \gamma}{\beta + \gamma} > 0
\] (4.9)

If this condition is not met, the social optimum must be a fully separated equilibrium, such that there exists a \(t \in [-1, 1]\) for which demand is zero. In that case, the optimal fares are still equal to zero, so any \(\{t_1, t_2\}\) that satisfies \(d(-1) = d(t^*) = d(1) = 0\) is an equilibrium, as social welfare does not depend on the exact departure times, and therewith the exact desired arrival times served. Shifting to an earlier time will allow more passengers with an earlier desired departure time to travel, but keeps exactly the same number of passengers with later desired departure times from traveling, such that the net effect on welfare is zero.

4.4 Full market separation

If the market is served by two private operators, two types of equilibria may emerge: either the two operators serve fully separated markets, such that they do not compete for a marginal passenger, or the market is fully covered. We will consider these two types of equilibria separately, starting with full market separation.

It may be optimal for two operators to set their fares and departure times such that \(d(-1) = d(t^*) = d(1) = 0\), so that their markets are fully separated and each operator acts as a monopolist on its own segment. In this case, there exists a desired departure time that
4.4 Full market separation

is so far away from both services that nobody with this desired departure time travels. To examine when this would happen, consider a single monopolistic operator who can set any departure time and fare, and faces the linear demand function in Eq. 4.1 for all \( t \in \mathbb{R} \), such that it is not constrained by a fixed time period.

Solving Eq. 4.1 to obtain the passengers with the earliest and latest desired departure times that are traveling in this situation gives

\[
\{t, \bar{t}\} = \left\{ t_1 - \frac{a - bf_1}{\gamma b}, t_1 + \frac{a - bf_1}{\beta b} \right\}
\]

(4.10)

This operator’s profits are then

\[
\pi_1 = \int_{t}^{\bar{t}} d(t)dt - F = f_1\frac{(\beta + \gamma)(a - bf_1)^2}{2\beta \gamma b} - F
\]

(4.11)

Naturally, these profits do not depend on the operator’s departure time choice, as long as both are within the \([-1, 1]\) interval, since there is now no unique fixed time period. The operator’s profit in Eq. 4.11 is maximized when

\[
f_1 = a/(3b)
\]

(4.12)

Substituting this back in Eq. 4.10 gives the passengers with the earliest and latest desired departure times, as a function of the parameters:

\[
\{t, \bar{t}\} = \left\{ t_1 - \frac{2a}{3\gamma b}, t_1 + \frac{2a}{3\beta b} \right\}
\]

(4.13)

Only if \( \bar{t} - t \leq 1 \) can two fully separated monopolists with fares as in Eq. 4.12 operate between \( t = -1 \) and \( t = 1 \). This implies that the equilibrium is fully separated if

\[
\frac{2(\beta + \gamma)a}{3\beta \gamma b} \leq 1
\]

(4.14)

since, in that case, a fully separated equilibrium always results in higher profits than an equilibrium with a covered market. If this condition does not hold, the market is fully covered. Hence, a fully separated equilibrium is more likely to occur when the maximum number of passengers (\( a \)), for a given \( t \), is smaller, when the demand sensitivity (\( b \)) is higher\(^4\), for higher costs of schedule delay, and for a larger difference between the cost of schedule delay late and the cost of schedule delay earlier. However, operators can only recover their costs if \( \pi_i \geq 0 \). Using Eq. 4.10, this implies that

\[
\frac{2(\beta + \gamma)a^3}{27\beta \gamma b^2} \geq F
\]

(4.15)

This, conversely, is less likely to occur when the maximum number of passengers is smaller, the demand sensitivity is higher, and for higher costs of schedule delay. In the absence of any subsidies, Eqs. 4.14 and 4.15 can only hold simultaneously if \( F \leq a^2/(9b) \).

\(^4\)Note that the effect of \( b \) runs via its impact on demand, given \( a \). In particular, for a given price and with equal \( a \), the demand elasticity is independent of \( b \).
4 A Hotelling model with price-sensitive demand and asymmetric distance costs

4.5 Equilibria with covered markets

Having derived when a separated equilibrium occurs, we can now examine the various possibilities for an equilibrium in which the market is entirely covered, such that the two operators compete for the marginal customer. If Eq. 4.14 does not hold, $D(-1), D(t^*), D(1) > 0$, since it would be suboptimal to stay in a situation where $D_1(-1)$ or $D_2(1)$ equal zero while $D(t^*) > 0$, and vice versa. The resulting equilibrium is considerably more complicated than the fully separated one, which is why we will start by assuming that $\beta = \gamma$, in order to derive tractable results, before we consider asymmetric schedule delay costs.

If markets are covered, the equilibrium fares and departure times depend on the order in which they are chosen if a sequential game structure is allowed. We will examine three possibilities: either fares and departure times are chosen simultaneously, or departure times are chosen first, while fares are chosen only after the departure times have been fixed, or vice versa. In all cases, we initially assume Nash behavior, moving to Stackelberg games in section 4.6.

4.5.1 Simultaneous departure time and fare choice

Puu (2002) analyses a Hotelling game in which locations and prices are chosen simultaneously, and derives an equilibrium where both suppliers charge equal prices, and $t_1 = -t_2$. However, as Sanner (2005) shows, this equilibrium only appears to be stable because of a calculation error. In reality, each operator could obtain a higher profit by choosing the same location as its competitor, while undercutting its competitor's price with an arbitrarily small amount. It would then serve the entire market of its competitor, plus at least part of its own original market. Hence, in this situation, no stable equilibrium exists.

This result continues to hold if there are more than two competitors or if demand functions are nonlinear, for the same reason; it is always possible for one competitor to take its direct neighbor’s place and undercut its price by an arbitrarily small amount, and then obtain the full market. As long as the two operators were competing for the marginal customer, the profit of the undercutting competitor will then increase. Only if one of the operators sets its fares and locations before the other can an equilibrium exist; we will briefly examine this game in section 4.6 below.

4.5.2 Fares chosen before departure times

We can find the equilibrium by backward induction, by first solving $\frac{\partial \pi_i}{\partial t_i} = 0$ for $i = \{1, 2\}$ to obtain the optimal departure times, substituting these in the operators’ profit functions, and then solving for the optimal fares. The optimal timing response functions for both operators are

$$t_1 = -\frac{4}{5} + \frac{1}{5\beta} \left( \frac{2a}{b} - 3f_1 + f_2 \right) + \frac{t_2}{5}$$

$$t_2 = \frac{4}{5} - \frac{1}{5\beta} \left( \frac{2a}{b} + f_1 - 3f_2 \right) + \frac{t_1}{5}$$

(4.16)
4.5 Equilibria with covered markets

Solving Eq. 4.16 to obtain the equilibrium departure times $t^*_i(f_1, f_2)$ gives

\[
\begin{align*}
    t^*_1 &= -\frac{2}{3} + \frac{1}{3\beta} \left( \frac{a}{b} - 2f_1 + f_2 \right) \\
    t^*_2 &= \frac{2}{3} - \frac{1}{3\beta} \left( \frac{a}{b} + f_1 - 2f_2 \right)
\end{align*}
\]

(4.17)

The equilibrium fares can then be obtained by solving $\frac{\partial \pi}{\partial f_i}\mid_{\{t_1 = t_1^*, t_2 = t_2^*\}} = 0$, which gives optimal response functions

\[
f_i = \frac{16a}{39b} + \frac{10}{39}(f_j + \beta) - \frac{1}{39} \sqrt{334 \left( \frac{a}{b} \right)^2 - 460(f_j + \beta)\frac{a}{b} + 295(f_j + \beta)^2}
\]

(4.18)

These derivations are tedious, and the resulting expressions have no intuitive interpretation.

The optimal response functions can be solved to obtain the equilibrium fares. However, this equilibrium is not stable. Substituting the equilibrium fares into the cross-partial derivatives of Eq. 4.18, $\partial f_i/\partial f_j$, results in a tedious expression which, however, is smaller than one for any positive $\beta$. This means that undercutting strategies are profitable as long as the operators are making positive profits. Instead of setting a fare equal to Eq. 4.18, any of the two competitors could set its fare an arbitrarily small amount lower. The other would then also adjust its fare, but by a smaller amount, since $\partial f_i/\partial f_j < 1$. In the timing subgame, this competitor could then simply choose the other’s departure time; with its lower fare, it would get the entire market. Since both competitors can use this undercutting strategy profitably as long as positive profits are made, the equilibrium is never stable.

This result continues to hold if the value of schedule delay early is higher than the value of schedule delay late. In that case, equilibrium profits are likely to be asymmetric, so undercutting may only be a profitable strategy for one of the operators, but this still results in instability. The same is true if demand or schedule delay function are non-linear; as long as $\partial f_i/\partial f_j < 1$ for at least one of the operators, and as long as the operators compete for the marginal traveler, any equilibrium where positive profits are made is unstable.

4.5.3 Departure times chosen before fares

Symmetric schedule delay costs ($\beta = \gamma$)

Again, we find the equilibrium by backward induction. Solving $\frac{\partial \pi_i}{\partial f_i} = 0$ for $i = \{1, 2\}$ gives the optimal fare response functions for both operators. Solving these gives the fare equilibrium:

\[
f^*_1 = \frac{a}{2\beta} + \beta \left( 2 + \frac{5}{4} t_1 + \frac{3}{4} t_2 \right) - \frac{1}{4b} \sqrt{4 \left( \frac{a}{b} \right)^2 + \frac{4\beta a}{b}(t_1 - t_2) + \beta^2 \left( 80 + 112t_1 + 45t_1^2 + 48t_2 + 22t_1t_2 + 13t_2^2 \right)}
\]

(4.19)

As well as one other root, which corresponds to the minimum profit.
Again, these derivations are tedious, and the resulting equations have no straightforward intuitive interpretation. Substituting them back in the original profit functions and maximizing each operator’s profit with respect to its departure times gives the equilibrium departure times. These do not have a closed form, and can only be evaluated numerically, which we will do below. What is important to note here is that in this game undercutting is not a profitable strategy. Rather than an arbitrarily small deviation from the first-stage subgame equilibrium, the only potentially successful undercutting strategy now requires an operator to take its competitor’s place; a major deviation. However, in the pricing stage, the only possible equilibrium is then the Bertrand equilibrium where both operators make zero profits. This is a less attractive option than the differentiated equilibrium above, and will therefore not be chosen.

We can also establish the interval in which an equilibrium exists. By construction, \(-1 \leq t_1 \leq t_2 \leq 1\). In this game, these conditions are met only when \(\beta \geq \frac{6a}{31b}\). For a smaller \(\beta\), there is no equilibrium. For \(\beta \geq \frac{4a}{31b}\), the equilibrium is separated. Using these bounds, it is also possible to calculate the range of \(\{t^*_1, t^*_2\}\):

\[
\lim_{\beta \to \frac{6a}{31b}} t^*_1 = -\frac{1}{2}, \quad \lim_{\beta \to \frac{4a}{31b}} t^*_1 = \frac{1}{2}
\]  

(4.21)

So, the two services are closer together than socially optimal. This is an important difference from many other horizontal differentiation models, notably that of d’Aspremont et al. (1979), in which competitors locate as far apart from each other as possible, such that they can exert local market power. The reason that this does not happen here is that, in our model, demand is price-sensitive. If one operator schedules its service further from the other, this does indeed decrease competition at the inner market boundary, allowing it to increase its fares in the second stage, as Eqs. 4.19–4.20 show. However, by doing so, it will also lose costumers with a desired departure time between the two services since, for these travelers, both schedule delay costs and fares have increased. Of course, this will also shift the inner market boundary. Hence, each operator’s departure time must be closer to the inner market boundary than to the closest outer market boundary, precisely because the latter is fixed, while the former moves in the same direction as a change in one operator’s departure time. How close it must be exactly depends on the optimal fare, and hence, on the value of schedule delay. When \(\beta\) approaches \(6a/(31b)\), the optimal two departure times approach 0, and when \(\beta\) is even smaller, the two operators will continuously swap places; no stable equilibrium emerges. When \(\beta\) is large, however, market areas are small, so the incentive to try and steal a competitor’s customers is smaller, and hence, the departure times are set further apart.

Fig. 4.1 shows the equilibrium fares, profits, departure times and social welfare relative to the optimum, for the range of values of schedule delay where an equilibrium exists. As already indicated by Eq. 4.21, departure times move further apart if the value of schedule delay increases. Fares and profits, however, are non-monotonic in \(\beta\). This is because an increase in the value of schedule delay has two effects. Firstly, an increase in \(\beta\) directly decreases the number of travelers, for any set of fares and departure times, as travel costs for all commuters increase. This will reduce optimal fares and profits. However, if travel costs

\[
f^*_2 = \frac{a}{2\beta} + \beta \left(2 - \frac{3}{4} t_1 - \frac{5}{4} t_2\right) - \frac{1}{4b}
\]  

\[
\sqrt{4 \left(\frac{a}{b}\right)^2 \beta^2 + \frac{4\beta a}{b} (t_1 - t_2) + \beta^2 \left(80 - 112t_2 + 45t_2^2 - 48t_1 + 22t_1t_2 + 13t_1^2\right)}
\]  

(4.20)
4.5 Equilibria with covered markets

for all commuters increase, they also increase for the marginal commuters, who have a desired departure time \( t = t^* \). Hence, there will also be fewer marginal commuters, which will reduce competition. This allows operators to increase their fares and profits. As Fig. 4.1 shows, this competitive effect dominates for smaller values of schedule delay. For larger values, the optimal departure times are already so far apart that the demand effect is stronger.

Social welfare, relative to the optimum, always decreases in the value of schedule delay. For low values of schedule delay, this is because fares are increasing in \( \beta \) and thus moving away from the optimum; although the departure times are moving closer to the optimum, this is less important, given that the values of schedule delay are relatively low. For higher values of schedule delay fares start decreasing slightly, but deviations from the optimal departure time are now so costly that although the departure times are moving towards the optimum, they are moving too slowly to offset the negative effect of an increase in \( \beta \) on welfare.

Asymmetric schedule delay functions \((\gamma > \beta)\)

If schedule delay functions are not symmetric in each commuter’s desired departure time, operators in the resulting equilibrium will charge different fares, and their departure times will not be at equal distances from zero. This complicates the analysis and, hence, this situation can only be evaluated numerically. Fig. 4.2 shows, for \( \beta = 5 \) and \( a/b = 10 \), the fares, departure times, operator profits and social welfare for a range of \( \gamma \). Although, naturally, the exact functions are specific to these particular parameters, other parameters result in very similar figures. Moreover, all variables only depend on the ratio between \( \beta \) and \( a/b \); not on the individual levels of these parameters.
Naturally, if the value of schedule delay late increases relative to the value of schedule delay early, both departures will move to an earlier time, such that fewer commuters are late. In an effort to gain the largest market share, both do so at a faster rate than the socially optimal departure times $t_{S_i}$. Hence, the first operator’s departure time initially moves closer to the optimum, while the second operator’s departure time, which is already earlier than optimal, continuously moves away from the optimum.

For moderate deviations of $\gamma$ from $\beta$, this allows both operators to increase their fares. However, the first operator’s market size decreases as it is squeezed towards its outer market boundary, and this reduces its profits. For large increases in $\gamma$, even the second operator loses, as demand for its service decreases too fast to be offset by its favorable position. Social welfare decreases in $\gamma$, relative to the first-best, as a result of higher prices, a less optimal departure time of the second operator and of course, in the same way as an increase in $\beta$, simply because suboptimal departure times become more costly.

### 4.6 Other games

As we have seen, games with simultaneous fare and timing choices, and games where fares are chosen first, never have pure strategy Nash equilibria. In games where departure times are chosen first equilibria only exist for limited ranges of parameters. It is therefore worth investigating which other game structures could result in equilibria where the above games fail. For the sake of brevity and simplicity, will limit our attention to situations with symmetric schedule delay cost functions, although, like before, it is possible to include asymmetries.
4.6 Other games

4.6.1 Stackelberg games

Stackelberg games, in which one operator sets its fare, departure time, or both before the other operator, may be a realistic representation of some real-world transport markets. A large operator, which is active not just in one market but operates many routes may, for example, have to decide on its fares and departure times much earlier than a small, flexible operator that only participates in one market. In this case, the operator that publishes its decisions first can choose them in such a way that it cannot be profitably undercut by the second operator. We will examine the Stackelberg equivalents of the three Nash games above: one situation in which the first operator sets its fare and departure time before the other, one in which the first operator sets its departure time before the other, followed by a separate second stage in which the operators set their fares in the same order, and the reverse, one in which the first operator sets its fare before the other, followed by a sequential departure time choice.

Fares and departure times set simultaneously per firm, and sequentially between firms

If the first operator decides on both its fare and its departure time before the other, and can commit to these decisions, it will choose them such that undercutting is not a profitable strategy for the second operator. This does mean that it has to accept a lower profit than it would get in some of the other games. Starting with the second stage, the second operator’s optimal fares and departure times \( \{t^*_2(t_1, f_1), f^*_2(t_1, f_1)\} \) can be found by simply setting \( \frac{\partial \pi_2}{\partial t_2} = \frac{\partial \pi_2}{\partial f_2} = 0 \). The first operator then maximizes its own profits subject to not only \( \{t^*_2, f^*_2\} \), but also another constraint, which specifies that the second operator’s profit must be greater or equal to the profit it would get if it took the first operator’s departure time, and set its fare an arbitrarily small amount lower:

\[
\pi^*_2 \geq \int_{-1}^{\tilde{t}} a - b(f_1 + Max[\gamma(t_1 - t), 0] + Max[\beta(t - t_1), 0])dt - F \tag{4.22}
\]

where \( \pi^*_2 \) is the send operator’s equilibrium profit, and \( \tilde{t} = Min \left[ t_1 + \frac{a - bf_1}{\beta}, 1 \right] \). Since this constraint will be binding for any set of parameters, Eq. 4.22 can be solved as a strict equality and used to substitute out one of the first operator’s decision variables. Since the constraint is nonlinear, and the resulting equilibrium will have asymmetric departure times and unequal fares, this game can only be solved numerically. It does have a unique pure strategy equilibrium for a large parameter space. Fig. 4.3 shows the equilibrium fares, profits, departure times and relative welfare for a range of \( \beta \).

As expected, the second mover in this game has an advantage, since the first mover has to choose a position that can not be undercut profitably. Hence, the first operator’s departure time is close to the outer market boundary, and its price is lower than in the corresponding Nash game. Naturally, it is therefore optimal for the second operator to set an earlier departure time and higher fare than it would do in a Nash game. As the value of schedule delay increases, resulting in a lower travel demand, the first operator can choose a more favorable position, as undercutting becomes less profitable. For very high values of schedule delay, the second-mover advantage all but disappears, and the equilibrium locations approach those of the Nash game.
4 A Hotelling model with price-sensitive demand and asymmetric distance costs

This game is similar to the previous, but has four separate stages; $f_1$, $f_2$, $t_1$ and $t_2$ are chosen in that order. Again, the resulting equilibrium is asymmetric; the second operator has a second-mover advantage, so $\pi_1^* < \pi_2^*$. Undercutting is still possible, and can still be profitable, especially for small values of schedule delay, but the first operator can again avoid this. It can be shown numerically that this game has an equilibrium for at least some values of $\beta$.

Departure times chosen before fares

The equilibrium of this game, in which $t_1$, $t_2$, $f_1$ and $f_2$ are chosen in that order, is the most tedious to compute, as only the last subgame has a closed form. However, numerical simulations show that, for all $\beta \geq 0.32a/b$, a pure-strategy equilibrium exists; this is a considerably smaller space than in the Nash game in which both operators set their departure times at the same time, followed by a simultaneous fare decision.

4.6.2 Regulation

As we have seen, operators can use their departure times as strategic instruments, and hence, regulation may be desired. Even if the first-best solution, in which fares are zero and departure times given by Eq. 4.8, is not feasible, it may be possible for the regulator to set either fares or departure times. The operators would then be free to set the remaining variables afterwards. This requires that regulators can commit to the choices they announce; operators must be
4.6 Other games

convinced that the announced fares or departure times will not be changed after they have announced their own decisions. If, for any reason, the regulator cannot commit, the game collapses from a two-stage game to a simultaneous game, in which the regulator and the operators effectively set fares and departure times simultaneously.

Hence, at least four different games should be considered: two sequential games, in which regulators set either fares ($F_{ST}$, where the subscript $S$ denotes a second-best social optimum) or departure times ($T_{SF}$), and two simultaneous games. Fig. 4.4 shows the performance of each game, relative to the first-best social optimum, for the range of $\beta$ in which each game has an equilibrium.

![Figure 4.4: Relative welfare of regulatory games](image)

Interestingly, all four games have stable equilibria over a large range of values of schedule delay, while only the first, in which the regulator sets the departure times, followed by a competitive pricing stage, has a stable equivalent in an unregulated market. In reality, completely unregulated transport markets are a rare occurrence, so this may explain why we usually observe stable equilibria. Indeed, there is some evidence that deregulation can lead to service instability (see, for example, Douglas, 1987). However, even the regulatory games do not have equilibria for all possible values of schedule delay. In particular, the games in which the regulator sets fares are unstable for small values of $\beta$.

As Fig. 4.4 shows, regulators have a first-mover advantage, regardless of which variable they are controlling; welfare in each sequential games is always higher than welfare in the corresponding simultaneous game. However, this advantage is relatively small for most parameters. A game in which departure times are chosen by the regulator is preferred, from a social perspective, if values of schedule delay are low. Conversely, when values of schedule delay are high, a game in which the regulator chooses fares performs better. This may sound counterintuitive since, when values of schedule delay are low, suboptimal departure times are relatively unimportant compared to suboptimal fares, so one would expect a more efficient outcome if the regulator set the latter. This does not happen because, for low values of schedule delay, the private operators set low fares anyway; it is better for the regulator to set the departure times, even if they are relatively unimportant. If values of schedule delay
are high, on the other hand, the operators would exercise their increased market power and raise fares far above the optimum; it would then be better for the regulator to set the fares instead, even though suboptimal departure times are also relatively costly.

Of the four games in Fig. 4.4, only the first, in which the regulator sets the departure times, followed by a competitive pricing stage, has a stable equivalent in an unregulated market. A comparison of the relative welfare in this game to the bottom right-hand panel in Fig. 4.1 shows that the increase in social welfare that results from regulatory intervention is relatively modest, especially for small values of schedule delay. Only when $\beta$ is very high, such that small deviations from the optimal departure times have a large effect on social welfare, is the regulatory game much more efficient.

4.7 Discussion

Naturally, our quantitative results depend on the assumptions we have made. One obvious way to generalize the models described above further would be to allow for non-uniform desired departure time distributions (see, for example, Tabuchi and Thisse, 1995; Janssen et al., 2005). This could, for instance, be achieved if Eq. 4.1 was replaced by

$$d(t) = \text{Max} [a(t) - bp(t), 0]$$  \hspace{1cm} (4.23)

Naturally, the integration in Eqs. 4.6–4.7 would be much more complex, though not necessarily impossible, depending on the exact functional form of $a(t)$.

Depending on the chosen distribution, this generalization would change the relative importance of the inner and outer market boundaries. If, for instance, the distribution was triangular, such that the largest number of passengers preferred a departure time somewhere in the middle of the time interval (for example, $a(t) = a_0 - |x|$), this would increase competition at the inner market boundary, as there would be more potential passengers there. Consequently, we would expect operators to move closer together than they do in the models above. This increased competition at the inner market boundary would also make the parameter space in which a stable equilibrium emerges smaller; in that regard, it could be expected to have a similar effect as a decrease in $\beta$ in section 4.5.3.

The addition of more competitors (see, for example, Brenner, 2005), capacity constraints (as in Wauthy, 1996a), nonlinearities, or crowding costs will, of course, also affect the results. However, the qualitative results, and the mechanisms of competition behind them, will remain the same. Circular Salop models, in which the time interval does not have endpoints, are much more difficult to use together with price-sensitive demand (see Gu and Wenzel, 2012, where demand is independent of the distance costs; a more general formulation would be even more complicated!), since they have two inner market boundaries, where the operators compete for the marginal customer.

Our models are, themselves, generalizations of earlier Hotelling models. The inclusion of price-sensitive demand and asymmetric distance costs, although adding realism, do create more complexity. One major drawback of our approach is that closed-form solutions can not always be obtained, and if they can, their interpretation is difficult. This is particularly true for the models with asymmetric distance costs. These asymmetries are important in some markets, such as transport, but in others, this generalization will unnecessarily complicate analyses. Price-sensitive demand also makes the results less intuitive, but, contrary to the asymmetric distance costs, it helps establish equilibria in situations where models with fixed
demand are unstable. Since the inclusion of price-sensitive demand is less arbitrary than some other measures to prevent instability (such as particular nonlinear formulations for the distance cost function), it might still be an attractive modeling choice.

Naturally, real-world transportation problems are more complex than the stylized models we have presented. In many settings, departure times will not only be driven by link-based competition, but also by the arrival and departure of connecting services. Hence, services can be complements as well as substitutes. It is, however, still important to first understand what happens on single links, especially since models that include connecting services will need to make other restrictive assumptions to remain tractable.

4.8 Conclusions

We have proposed a methodology to model the scheduling decisions of competing transport operators, using a generalization of Hotelling’s horizontal differentiation model. This model, which includes price-sensitive demand, often has equilibria where other models are unstable. Equilibria in our model can be interior and do not necessarily result in minimum differentiation, and never in maximum differentiation. Indeed, the two competitors normally schedule their services closer together than optimal. This happens because, when demand is price-sensitive, operators have incentives to schedule their services closer to the inner market boundary than to the outer edges of the market, since the inner market boundary can be pushed in the direction of the competitor, while the outer edges of the market are fixed. Games where prices are chosen before or simultaneously with locations, on the other hand, have no stable Nash equilibria.

We have also shown that it is possible to include asymmetric schedule delay functions in our model. Asymmetric schedule delay functions generally lower the relative welfare of the game. Since asymmetric schedule delay functions result in asymmetric equilibria, they do make the calculation of the equilibria much more involved, and for some parameters cause unstability if undercutting is profitable.

A Stackelberg structure, in which one operator sets its decision variables before the other often helps to establish equilibria in games where they are not present with Nash behavior. In these games, the first mover can deliberately choose a position such that its competitor has no incentives to undercut; a consequence is that there is a second-mover advantage. Similarly, regulation can create equilibria in situations where an unregulated market fails to do so. If the socially optimal locations and prices are not attainable, regulating one of these two variables can result in a modest efficiency improvement; the value of schedule delay determines which of the two results in the greatest gain.

Although our quantitative results depend on our assumptions, the qualitative results, and the mechanisms of competition that drive them, are valid for a much larger class of models. Most importantly, the models proposed above show that departure times can be strategic instruments, and should therefore be of interest to regulators.
5 Stochastic user equilibrium traffic assignment with price-sensitive demand: do methods matter (much?)

5.1 Introduction

The previous chapters have discussed pricing, scheduling, and regulation in transport networks. In these chapters, user equilibria were always deterministic. As mentioned in the introduction, deterministic user equilibrium models have important disadvantages, particularly in larger network models. The analysis of transport infrastructure investment in Chapter 6 will therefore use stochastic user equilibrium (SUE) models. SUE traffic assignment models are used in a wide range of applications. Starting with Dial (1971), who proposed a simple logit model to analyze route choice in a network, many different formulations have been developed. These range from highly simplified, tractable models to very complex models that can only be solved with numerical methods. Most are static, although some dynamic formulations have also been proposed (Ben-Akiva et al., 1984, 1986). Choosing a SUE model for a specific application is not straightforward. In preparation of Chapter 6, the present chapter will compare several popular models, to determine which of them are best.

In a deterministic Wardropian equilibrium Wardrop (1952), the generalized prices (i.e. costs plus tolls) of all used routes are equal, and lower than those of all unused routes. This implicitly assumes that users have perfect knowledge about the costs of all routes, and that all relevant user attributes can be perfectly observed. SUE models instead use random utility discrete choice theory, which assumes that the utility users derive from a given route has a stochastic part, which cannot be explained by an observer, but which follows a known distribution. The resulting equilibrium differs from the Wardropian equilibrium in that the systematic (or, deterministic) generalized prices of all routes are typically not equal. Moreover, although the probability that a route will be used can approach 0 arbitrarily closely, it never reaches it. Besides being more realistic in many settings, this is also often computationally convenient.

In this chapter, we first consider two ‘workhorse’ user equilibrium models: a multinomial probit, and a logit model. Both models are used regularly in the recent literature (see e.g. for logit: Meng et al., 2004; Yang, 1999; Yang et al., 2001, and for probit: Connors et al., 2007; Uchida et al., 2007; Meng et al., 2012). Probit models can account for partially overlapping routes and routes with significantly different lengths, while simple logit models can not properly handle overlap, and assume that all route costs are subject to the same level of stochasticity. However, logit models have closed form solutions for choice probabilities, while probit equilibria can only be determined using sampling techniques or numerical integration.

---

1 This chapter is based on joint work with Vincent van den Berg and Erik Verhoef. An earlier version of this chapter has been published as a Tinbergen Institute Discussion Paper (van der Weijde et al., 2013a). I thank Paul Koster, Andrew Koh, and the participants of the 2013 INFORMS Annual Meeting for helpful comments.
and are therefore highly computationally intensive. It is therefore very useful to know how important the differences between these models are likely to be in real-world situations. Although these standard logit and probit models are widely used, several generalizations and extensions have been proposed. As an example of such an advanced logit model, we also include the generalized nested logit (GNL) model in our analysis. A relatively novel model, the GNL (Wen and Koppelman, 2001) can partially account for overlapping routes (though it does so in a different way than the probit model): for this reason, it has been proposed as a possible SUE assignment model (Bekhor and Prashker, 2002). Unlike the probit model, it cannot properly account for the fact that routes have different lengths; however, it does have closed-form probabilities. Naturally, there are many other SUE models, some of which have been developed more recently, and many of which are substantially more complex than the models used in this chapter (see e.g. Ben-Akiva and Bierlaire, 2003 for a review of some of those). However, as this chapter aims to analyze models that can be used as the lowest level of larger, theoretical economic models, we will not consider these.

Although many studies investigate the differences between competing discrete choice models, none of these are directly applicable to our setting. Many focus only on estimation or calibration, rather than simulation (e.g. Cascetta et al., 1996). For estimation, it is important that models fit existing data well. For simulation, the representation of one particular flow pattern is not of primary concern, as most models can be calibrated so as to achieve that. Instead, it is important to see what happens if we move away from the calibrated state, through, for instance, tolling. The different models imply different substitution patterns, so the effects of these changes could be different. This could have important implications for operational decisions and policies.

Others compare models in highly simplified settings; as we will see, their results do not necessarily carry over to more realistic situations (e.g. Florian and Fox, 1976; Daganzo and Sheffi, 1977; Prashker and Bekhor, 2004). Finally, some studies compare the very complex SUE models mentioned above (a good example of this last category is the extensive work by Ramming (2001)). Most studies in all of these categories compare models based only on route choice choice probabilities, or route flows. For an economic analysis, these variables are interesting, but often other variables may be more important. If, for instance, the models are used to derive optimal road prices, pure traffic flows are arguably less important than profits.

We therefore compare the different SUE models in a setting that has:

1. A network that is more representative of real-world situations than the simplest ‘toy’ networks that have been used in the past. Specifically, we model a network with several pairs of origins and destinations (ODs) and multiple overlapping routes between each OD-pair, which share links with routes that are used between other OD-pairs. Using this network, we can not only look at the effects of different SUE models on an OD-level, but also examine how these effects interact in a network.

2. Congestion; specifically, we use the realistic congestion function proposed by US Bureau of Public Roads (1964). As we will show in section 5.2.2, congestion can have a large effect on the differences between SUE models.

3. Price-sensitive demand, through the addition of an alternative ‘virtual’, uncongested route between each OD-pair, that does not overlap with any others.

4. Parameters that are calibrated correctly. If all assignment models are calibrated to the same traffic flows, as would happen in reality, this might give different results than a
situation where the assignment models have the same parameters, where this is possible (i.e., if the models have the same marginal utility of money and the same average variance; naturally, they will still have a different variance-covariance matrix). Therefore, we consider two cases: one in which the assignment models have equivalent parameters, and one in which the logit models are calibrated to the flows resulting from the probit model. We also investigate how the introduction of additional (alternative-specific) parameters, which can be used for calibration, affects the differences between the models.

To examine the differences between the assignment models, we test their performance in a 12-node network. In particular, we evaluate how the different assignment models affect policy decisions, such as profit-maximizing and welfare-maximizing tolls, and the effects of such policies. As we will show, the differences between models are, in many instances, small. However, improper calibration can lead to large differences, and significant reductions in welfare. This indicates that, in larger networks, where it is much more difficult to use probit-based assignment models, it may be justified to use simpler logit-based models.

The next section defines the assignment models, and explores the theoretical differences between them, using a very simple network model with only one origin and one destination. Section 5.3 gives an overview of the methodology we use for the more realistic simulations, the results of which are presented in section 5.4. Section 5.5 concludes.

5.2 Theory

5.2.1 SUE assignment models

All the stochastic user equilibrium assignment models we examine are based on random utility theory. In these models, the utility users derive from an alternative (which, in traffic assignment, can be a route through a network, the use of a specific link, or the option to stay home and not travel) consists of a deterministic, and a stochastic part:

\[ U_{ir} = V(-c_r) + \epsilon_{ir} \]  

where \( U_{ir} \) is the utility user \( i \) derives from alternative \( r \), which is made up of a deterministic value function \( V \) and a stochastic term \( \epsilon_{ir} \). For simplicity, we define the value as a function of the generalized price \( c_r \) of an alternative only; we assume that benefits are the same for all alternatives. Since we have only one user class, \( V \) is also independent of personal characteristics. Although, in much of the recent literature, this random term is assumed to capture the uncertainty that transport users face, it was originally intended to capture measurement errors made by the observer (reflecting, for instance, the fact that users differ in unobservable characteristics, which influence the utility they derive from the use of an alternative). Different assignment models assume different joint distributions for the random terms.

**Multinomial probit (MNP)**

The first discrete choice model we consider is one of the most flexible. In the multinomial probit (MNP) model, the random terms \( \epsilon \) follow a multivariate normal distribution: \( \epsilon \sim \mathcal{N}(0, \Sigma) \). Importantly, if \( \Sigma \neq I \), random terms are allowed to be correlated across all alternatives, and the variances can differ across alternatives. We can therefore set \( \Sigma \) such that the covariances
between different routes, \(0 \leq \sigma_{rt} < 1\), indicate to what extent routes overlap, and variances \(\sigma_r^2\) that vary with the length of each route.

As already stated, we will model the price-sensitivity of demand through the inclusion of a no-travel alternative, which has a given cost (or disutility). This is not the only way to make travel demand price-sensitive. It is also possible to make the total travel demand between an origin and a destination a direct function of the expected utility of the discrete choice. However, particularly in probit models, this is considerably more complicated, and for our purposes, it yields few advantages. Our approach still results in a downward-sloping demand curve, which is non-linear, and whose slope reflects by the elasticity of the no-travel alternative choice probability.

There are several ways to derive the covariances \(\sigma_{rr'}\) and variances \(\sigma_r^2\). The most straightforward way, which we will use, is to formulate link-based random terms \(\varepsilon_{il}\). The total random utility faced by a user who takes a given route \(r\), \(\varepsilon_{ir}\) is then the sum of all link-based random terms, over all links that constitute the route. Since the sum of two or more independent normal distributions is also a normal distribution, the covariance between two routes is then the sum of the variances of the links they share. We set link-based variances \(\sigma_l^2 = sK_l\), where \(K_l\) is the length of link \(l\), and \(s\) a parameter. For simplicity, we set the variance of the no-travel alternative to 1. The variance of route \(r\) is then simply \(\sum_l a_{lr} \sigma_l^2\), where \(a_{lr} = 1\) if link \(l\) is part of route \(r\), and zero otherwise. The covariance between routes \(r\) and \(r'\) is \(\sum_l a_{lr} a_{lr'} \sigma_l^2\); the covariance between any route and the no-travel alternative is zero. We use \(s\) to control the average route variance \(\bar{\sigma}^2\), which we normalize to 1. Finally, we set \(V(-c_r) = -c_r/10\), where the coefficient is chosen such that, in equilibrium, a significant number of routes is used. However, note that, in this model, we cannot differentiate between the average variance of the random terms and this coefficient. Setting \(\bar{\sigma}^2 = 10\) and \(V(-c_r) = -c_r\) would yield the same results.

MNP models do not have closed-form probabilities; instead, numerical methods are necessary to determine them. We use a Monte Carlo simulation, where random terms \(\varepsilon\) are sampled from the multivariate normal distribution described above. Based on these sampled random terms and the deterministic route costs, probabilities can be estimated. Naturally, since these probabilities depend on traffic volumes, we need an iterative procedure to find the user equilibrium assignment. We use the Method of Successive Averages (MSA), which consists of the following steps:

1. Determine initial link flows \(v_{l0}\) (e.g. estimate probabilities based on free-flow travel costs) and obtain link-based travel costs \(\chi_l\). Set \(n = 1\).

2. Calculate auxiliary link flows \(v_l^A\) based on costs \(\chi_l\).

3. Set \(v_{ln+1} = v_{ln} + n^{-1} \left(v_l^A - v_{ln}\right)\) and recalculate costs \(\chi_l\). Set \(n = n + 1\).

4. Repeat steps 2–3 until \(|v_{ln+1} - v_{ln}|\) becomes sufficiently small.

Since step 2 requires a Monte-Carlo simulation, and the number of iterations the MSA needs to converge can be high, this model can be computationally intensive. It can, however, account for both overlapping routes, and routes of different lengths, which the other models we will examine can not fully take into account.

Since we are interested only in the differences between the assignment models, and the MNP is the most natural benchmark, we will use the MNP flows to calibrate the other assignment models. Alternatively, one could obtain flows from some other, external model,
and calibrate all assignment models using those flows; this would not add anything to our analysis. This does imply that our results can only be used to analyze the differences between assignment models; it does not indicate which model is better at approximating some particular real-world dataset.

**Logit**

The simplest logit model assumes that all random terms $\varepsilon_{ir}$ are independent. However, the introduction of elastic demand through the addition of a no-travel alternative makes this assumption particularly unrealistic; it is logical to assume that routes more similar to each other than to the no-travel alternative. We therefore use a somewhat more advanced nested logit (NL) model (Ben-Akiva, 1974), in which one nest contains all the physical routes, and the other the no-travel alternative. In the NL model, the random terms follow a multi-variate extreme value distribution, and there is correlation within, but not between, nests. Hence, if we only consider the users that have already chosen to travel, the traffic assignment model is still a standard logit model; the NL formulation just allows us to model the demand elasticity correctly. To avoid confusion, however, we will refer to this model as an NL model in what follows.

The Generalized Extreme Value (GEV) function of this model (see McFadden, 1978) for a given OD-pair (or market) $m$ is

$$G_m \left( e^{V(-C)}, e^{V(-c_1)}, \ldots, e^{V(-c_R)} \right) = e^{V(-C)} + \left( \sum_r a_{rm} \left( e^{V(-c_r)} \right)^{1/\mu} \right)^{\mu}$$

(5.2)

where $a_{rm} = 1$ if route $r$ serves market $m$ and zero otherwise, $c_r$ is the cost of route $r$ and $C$ the cost incurred by a user who decides not to travel. $0 < \mu \leq 1$ is a parameter, indicating the dissimilarity between the option to not travel on the one side, and all possible routes on the other (more precisely, this implies a correlation of $(1 - \mu)^2$ between the random terms attributed to different routes (Daganzo and Kusnic, 1993). We use linear value functions $V(-C) = -\beta C$ and $V(-c_r) = -\beta c_r$.

As stated before, we will examine two versions of this model. In the first version, we set $\beta = \pi/\sqrt{10 \times 6}$ to achieve the same variance as MNP model: since logit models assume that the random terms follow a Gumbel distribution, which has a variance of $\pi^2/6$, and the MNP model defined above has an average variance $\tilde{\sigma}^2 = 1$ and $\beta = 1/10$, we have to correct for this in the NL model. We also set $\mu$ such that $(1 - \mu)^2$ is equal to the average overlap between routes. In the second version, we instead estimate $\beta$ and $\mu$, using Maximum Likelihood Estimation (MLE) on the route flows resulting from the MNP model; the resulting parameters are different than those in the first model.

Whatever the value of $\beta$, the resulting route choice probabilities have closed forms:

$$p_{mr} = \frac{\left( e^{-\beta c_r} \right)^{1/\mu} \left( \sum_r a_{rm} \left( e^{-\beta c_r} \right)^{1/\mu} \right)^{\mu}}{\sum_r a_{rm} \left( e^{-\beta c_r} \right)^{1/\mu} + e^{-\beta C}}$$

(5.3)

and consist of two parts. The first gives the probability that a randomly selected user chooses
route \( r \), given that that particular traveler has already decided to travel. The second part gives the probability that this user travels at all. Note that, in this case, \( \sum_r P_{mr} < 1 \), since some users will choose not to travel.

Although this NL model allows for correlation within nests, and can therefore control for the difference between traveling and not traveling, it still assumes that, given the choice of a nest, the alternatives are independent. Hence, contrary to the MNP, overlapping routes are not properly accounted for. Moreover, the variance is the same for all alternatives, so it is also impossible to properly account for the fact that routes have different lengths.

**Nested logit with alternative-specific constants (NL-ASC)**

To see how better calibration affects the difference between models, we also examine a version of the NL model that has additional parameters, which can be used for calibration. We do this by defining an alternative-specific constant (ASC) \( A_r \) for every route (not including the no-travel option), such that the value functions become:

\[
V(-c_r) = A_r - \beta c_r
\]  

Using MLE, these constants are then estimated, together with the \( \beta \) and \( \mu \). There is no constant for the no-travel alternative, to avoid overspecification.\(^3\)

Naturally, the flow pattern resulting from this model will be closer to the MNP flows in the calibrated state, as there are many additional parameters to use for calibration. However, as the MNP model does not have alternative-specific constants, the NL-ASC model may perform worse when we move away from the calibrated state, for instance, when a link in the network is tolled. The direction and magnitude of this effect depends on how the introduction of these ASCs changes the calibrated values of \( \beta \) and \( \mu \).

**Generalized nested logit (GNL)**

Like the MNP model, the GNL model (Wen and Koppelman, 2001; see Daly and Bierlaire, 2006 for an alternative formulation), also sometimes called the cross-nested logit model, can control for overlap in routes. In the GNL, alternatives can belong to several nests; inclusion parameters \( \alpha_{rl} \) indicate which share of alternative \( r \) belongs to nest \( l \). In a traffic assignment model, each link would be a nest (with an additional nest containing the option not to travel), and the share of a route that uses a specific link can be approximated using free-flow travel speeds (Bekhor and Prashker, 2002). The GEV function for market \( m \) is then

\[
G_m \left( e^{V(-C)}, e^{V(-c_1)}, ..., e^{V(-c_R)} \right) = e^{V(-C)} + \sum_l \left( \sum_r \alpha_{rm} \left( \frac{1}{\alpha_{rl}} \right) \right)^{\frac{1}{\mu}} \]  

where

\[
\alpha_{rl} = \frac{\chi_l}{c_r |_{n_l = l}}
\]  

\(^3\)The inclusion of alternative-specific constants is an often-used way to account for differences between routes that are not easily observed. The number of constant is usually more limited, for computational reasons. In our context, it is possible to include an ASC for all but one alternative; this will give the largest number of calibration parameters possible without overidentification. Hence, this is a useful benchmark; a model with fewer ASCs will, likely, give results that are somewhere in between our NL model, and our NL-ASC model.
and $\chi_l$ if the cost of using link $l$. Again, we define $V(-C) = -\beta C$ and $V(-c_r) = -\beta c_r$. Probabilities still have a closed form, and are given by

$$P_{mr} = \sum_l \left( \frac{\alpha_{rl} e^{-\beta c_r}}{\sum_r \alpha_{rl} m} \right)^\mu \left( \frac{\sum_r a_{rtm}}{\sum_l \sum_r a_{rtm}} \right)^\mu \mu_l$$  \hspace{1cm} (5.7)

Note that, although this expression is very similar to Eq. 5.3, the crucial difference is that there are now no longer two nests per market (travel and no travel); instead the number of nests is equal to the number of links plus one. Again, Eq. 5.7 only gives the route choice probabilities; there is also the option to not travel, so $\sum_r P_{mr} < 1$. Bekhor and Prashker (2002) propose to directly relate $\mu_l$ to the network topology, by making it an inverse function of the average inclusion coefficient of routes using link $l$. In our setting, all links share the exact same characteristics, so we also average over all links to get a single $\mu = \sum_l (1/L) (1 - (1/R_l) \sum_r \alpha_{rl})$, where $L$ is the total number of links, and $R_l$ the number of routes passing through link. We then set $\mu_l = \mu \forall l$, while $\beta$ has the same value as in the parameter-equivalent NL model. As with the NL model, we also examine a calibrated version, in which we use MLE to find a $\beta$ and $\mu$ that provides the best fit with the probit route flows.

Although GNL models have closed-form probabilities, which is very convenient for a traffic assignment model, the implied covariances between alternatives do not, and are not easy to calculate. They are given by

$$\text{Cov}(\varepsilon_r, \varepsilon_{r'}) = \int_{-\infty}^{\infty} (F(\varepsilon_r, \varepsilon_{r'}) - F(\varepsilon_r) F(\varepsilon_{r'})) \, d\varepsilon_r d\varepsilon_{r'}$$  \hspace{1cm} (5.8)

where

$$F(\varepsilon_r) = \exp(-\exp(-\beta \varepsilon_r))$$  \hspace{1cm} (5.9)

and

$$F(\varepsilon_r, \varepsilon_{r'}) = \exp\left(- \sum_l \left( \frac{\alpha_{rl} e^{-\beta \varepsilon_r}}{\mu_l} + \frac{\alpha_{r'l} e^{-\beta \varepsilon_{r'}}}{\mu_l} \right) \right)$$  \hspace{1cm} (5.10)

(Marzano and Papola, 2008; Lemp et al., 2010). These covariances are different than those of the MNP model, but do capture some of the overlap between routes.

### 5.2.2 Model differences

Before considering a more complex setting, it is useful to look at the simplest possible network in which the differences between the models above can be illustrated. This simple network, which is often used in the early literature on discrete choice models (e.g. Florian and Fox, 1976; Daganzo and Sheffi, 1977) has three routes, of which two partially overlap. Fig. 5.1 gives a graphical representation of such a network, where all routes between A and C have a length of 1, and the two routes that pass through B share a length of $1-x$. We will have an independent route 1; the others routes are denoted 2 and 3. For simplicity we assume, for a moment, that the no-travel alternative is not available, such that the NL model collapses to the simplest possible multinomial logit model.

If there is no congestion, and hence, all link costs are constant and equal, the logit model will assign 1/3 of the total flow to each of the routes, which, arguably, is unrealistic. A probit
model can account for the overlap, by setting

\[
\sum = \begin{bmatrix}
\sigma^2 & 0 & 0 \\
0 & \sigma^2 & (1-x)\sigma^2 \\
0 & (1-x)\sigma^2 & \alpha^2
\end{bmatrix}
\]

(5.11)

The probabilities of the two overlapping routes will then increase in \(x\), and the probability of the independent route will decrease, as shown in the left-hand panel of Fig. 5.2. Daganzo and Sheffi (1977) show a similar figure, and assuming that the MNP model is the correct one, argue that the logit model is often highly biased, because it fails to take the correlations between routes into account.

However, in traffic assignment models, links are usually congestible, so the costs of each route depend on the fraction of travelers that uses it. In our simple model, we can introduce congestion by, for instance, making link costs per unit of distance a linear function of link flows. The user costs are then

\[
c = \begin{bmatrix}
1 + 3p_1 \\
1 + 3(xp_2 + (1-x)(p_2 + p_3)) \\
1 + 3(xp_3 + (1-x)(p_2 + p_3))
\end{bmatrix}
\]

(5.12)

where \(p_l\) is the probability, or fraction of the total flow assigned to route \(l\), and the gradient of the congestion cost function is set at 3 to generate a realistic fraction of congestion costs to total costs. The right-hand panel of Fig. 5.2 shows the fraction of the total flow assigned to route 1 that results from these route costs.

As the right-hand side of Fig. 5.2 shows, the introduction of congestion significantly reduces the difference between the logit and probit models. This happens because, when the links are congestible, users of the two overlapping routes impose a congestion externality on each other, since they both use link BC. This makes routes 2 and 3 less attractive, especially if \(x\) is small. The random terms in the logit model are still independent, but the systematic utilities \(V(-c_r)\) now share a common term. Hence, the difference between the logit and probit models becomes less important. It does not completely disappear, however, and is still significant if the amount of overlap between routes 2 and 3 is big.

So far, we have only compared the simplest possible logit model with the probit model. The left-hand panel in Fig. 5.3 shows the results of a similar exercise, without congestion,
5.2 Theory

Figure 5.2: Logit vs. probit

Figure 5.3: GNL
this time including the GNL model, with

\[
\alpha = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & x & 0 & 1 - x \\
0 & 0 & x & 1 - x
\end{bmatrix}
\]

and \(\mu = 2x/3\), which is chosen to generate route choice probabilities close to the MNP model. As Fig. 5.3 shows, the MNP and GNL fractions are very close, even in the model without congestion. The right-hand panel of Fig. 5.3 shows the covariance between the two overlapping routes in both models, which are calculated using Eqs. 5.8 and 5.11; they are also very close. Hence, the GNL model seems to approximate the MNP model well, at least in this simple example. Moreover, there is no clear relation between the amount of overlap and the difference between the two models.

The simple network in Fig. 5.1, has only one market; travel from A to C. In more realistic models, there are more complex network effects; links are used by travelers in multiple markets. Depending on the network structure, this can further reduce the difference between the assignment models. If, for instance, the top link in Fig. 5.1 is also part of route that serves another market, and that route partially overlaps another, the two differences between logit and probit probabilities may cancel each other out. Moreover, note that, for there to be a difference between the models, all three routes (two overlapping, and one separate) must not only exist, but also be used by a significant fraction of travelers. In a larger congested network, this is not necessarily the case; a large amount of congestion on a major link between two large nodes could, for instance, make that link so unattractive to travelers between other nodes, that all routes using it would be assigned very low choice probabilities.

The introduction of an additional alternative, representing the option not to travel, affects the models in different ways. On the one hand, the addition of a non-overlapping alternative in every market makes it more likely that a situation such as the one in Fig. 5.1 exists, and hence, that there are significant differences between assignment models. On the other hand, however, if that alternative is added as a separate nest in a nested logit model, as we will do, this gives an extra parameter (the dissimilarity between traveling and not traveling) which can be used to calibrate the logit model, and hence, reduce the difference with the probit model.

5.3 Simulation methodology

5.3.1 Network

To compare our three assignment models, we will apply them to the simple network shown in Fig. 5.4, where all links are bidirectional (and congestion levels are direction-specific), and the size of each node indicates the potential demand for travel. Nodes indicated with asterisks are connection nodes only. All links are 10 km long, and we consider all possible non-circular routes.\(^4\) This network layout allows us to examine partially overlapping routes, and interactions between markets. Importantly, there is not only overlap of potential but unused routes; in the base case equilibrium, which we will present below, there is still a significant amount of overlap if we only consider routes that are used by significant fractions

\(^4\)In larger models, the set of potential routes is usually restricted, to reduce the computational intensity of models; a large literature proposes methods to do this efficiently. Since this is a separate problem from the choice of the assignment model, we will not discuss it here.
5.3 Simulation methodology

(e.g., > 5%) of travelers.

All links in the network are congestible; we model congestion using the well-known Bureau of Public Roads (BPR) function (US Bureau of Public Roads, 1964), a widely used approximation of the congestion costs of highway travel:

$$\chi_l = vot \cdot \frac{K}{S_f} \left(1 + 0.15 \left(\frac{v_l}{V}\right)^4\right) + f_l$$  \hspace{1cm} (5.14)

where $vot$ is the value of time, $K$ the link length, $S_f$ the free-flow travel speed (here, all these parameters are assumed to be the same for all links), $f_l$ a potential toll (or fare), and $v_l/V$ the volume-capacity ratio on link $l$. We set the latter set such that the ratio of congestion costs to total costs remains within realistic limits; specifically, such that the term multiplying the value of time in Eq. 5.14 is between 2 and 12. The traffic volume on a link is the sum of the realized demand of every route using the link:

$$v_l = \sum_m \sum_r a_{rm} a_{lr} p_{mr} D_m$$  \hspace{1cm} (5.15)

In the logit models, route choice probabilities $p_{mr}$ and route flows $p_{mr} D_m$ can be calculated directly. In the probit model, this is more involved. Each iteration of the MSA uses new multivariate draws, which are independent of the previous one; hence, there is simulation noise in every iteration. Link flows are calculated as a weighted average of all auxiliary flows in previous iterations, which averages out the simulation noise. Route flows can be calculated in every iteration, but as they are not averaged, they will contain a much larger amount of noise.

The most straightforward way to accurately estimate route flows is to run one last iteration with a large sample size, after the link flows have converged, and calculate route flows in that iteration. This can be highly computationally intensive, especially in larger networks. Another solution, which we will use, is to calculate a successive average of route flows in all iterations; a similar procedure as the one followed to obtain link flows (except for the fact that averaged link flows are used in the following iterations, whereas route flows are just stored). This method uses information that is already available, and gives consistent results,
provided that the number of iterations is large enough, such that the remaining simulation
noise in the resulting average route flow is sufficiently small.

Finally, the cost of taking a given route is simply the sum of the the link-based costs over
all links that make up the route

\[ c_r = \sum_l a_{lr} \chi_l \]  

(5.16)

5.3.2 Demand

The potential demand (or demand function intercept) in market \( m \) is calculated with a simple
gravity model, such that it increases in the size \( N_i \) of each of the two nodes that form the
market, and decreases in the distance \( K_m \) between them:

\[ D_m = \prod_{i \in m} N_i \frac{\delta}{K_m} \]  

(5.17)

where \( \delta \) is a parameter. The realized demand depends on the generalized price of travel,
through the SUE models.

5.3.3 Welfare and profit

It is difficult to define a consistent welfare measure across all models. Logsums are a natural
choice for the logit models, but the lack of a closed form in the probit model makes calculating
an equivalent measure there more complicated. We therefore use the Rule of Half, which is
often used for policy analyses. It approximates the welfare gains from a certain policy policy
change (in our case, a change in toll) by

\[ \Delta W = \sum_r \frac{1}{2} \left( q_r^1 + q_r^2 \right) \left( c_r^1 - c_r^2 \right) + \left( \pi^2 - \pi^1 \right) \]  

(5.18)

where \( (\pi^2 - \pi^1) \) is the profit change resulting from the change in toll, \( c_r^1 \) and \( c_r^2 \) the route
costs before and after the change respectively, and \( q_r^1 \) and \( q_r^2 \) the numbers of users choosing
each route (this includes the no-travel alternative) before and after the change. Hence, it
approximates the demand curve between \( q_r^1 \) and \( q_r^2 \) with a linear function. To calculate the
welfare gains of a specific toll \( f \), we use the cumulative benefits of all changes, up to a toll of
\( f \). This is a significantly better approximation than simply using the Rule of Half once, where
the situation before the policy change would have a zero toll (see Nellthorp and Hyman, 2003).
In the latter case, the whole demand function between a zero toll and a toll of \( f \) would be
approximated with a linear function, which is particularly inappropriate in a discrete choice
setting, where demand functions are usually highly convex. Our approach uses a piecewise
linear approximation. The step size is, to some extent, arbitrary, but the decision to use just
a single step would be too; moreover, increasing the number of steps has a quickly decreasing
impact on the results.

Although this piecewise linearization is obviously a simplification, it performs well. Figure
5.5 shows, for the calibrated NL model, how the approximation compares to the expected
welfare when we vary the toll one of the links in the network; the difference between the two
is only substantial if the toll is so high that the flow on the tolled link approaches zero, which
is generally not a situation of interest. Since the expected welfare is much more difficult to
calculate in a probit model, we use the approximation in what follows.
5.4 Simulation results

Figure 5.5: Expected welfare and Rule of Half approximation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$</td>
<td>35</td>
</tr>
<tr>
<td>$vot$</td>
<td>10</td>
</tr>
<tr>
<td>$V$</td>
<td>15000</td>
</tr>
<tr>
<td>$K$</td>
<td>10</td>
</tr>
<tr>
<td>$S_f$</td>
<td>120</td>
</tr>
<tr>
<td>$N_i$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>225000 0 0</td>
</tr>
<tr>
<td></td>
<td>0 200000 0</td>
</tr>
<tr>
<td></td>
<td>0 0 275000</td>
</tr>
<tr>
<td></td>
<td>350000 0 0</td>
</tr>
<tr>
<td>$\delta$</td>
<td>50000</td>
</tr>
</tbody>
</table>

Table 5.1: Initial parameters

5.3.4 Parameters

Initial parameters are shown in Table 5.1. We initially also set all $f_l = 0$, and estimate the parameters in the calibrated versions of the logit models using the flows resulting from the probit model in that situation. We then increase one of the fares, and use the calibrated models to examine the effects of this change.

5.4 Simulation results

5.4.1 Parameter-equivalent models

We first compare link flows resulting from the three theoretically equivalent models: the MNP, NL with comparable coefficients, and GNL with comparable coefficients. Starting from a situation where no links are tolled, we increase the toll on the link between nodes 5 and 8, in both directions. Fig. 5.6 shows the resulting flows $v$, as a function of these tolls $f_{58} = f_{85}$, for four representative links. Flow patterns on the other links either do not vary significantly with this toll, or display similar patterns as those in Fig. 5.6.
Since these three models imply different substitution patterns, their resulting flows are, of course, not exactly the same. However, the differences between the models are relatively small, even on the link that is tolled (the top left panel in Fig. 5.6). As we have shown in section 5.2.2, this is, to a large extent, the result of congestion. Because all links are congested, the utilities users derive from routes always correlate if routes overlap, even in the NL model. The presence or absence of correlation between the stochastic parts of the utilities still has an impact, but not nearly as big as one might expect. Moreover, since most links are used in several markets, OD-level differences may partially cancel each other out on the network level.

Fig. 5.6 also shows that the GNL model, which allows for some correlation between random terms of routes, often results in flows closer to the Probit model, and it is never significantly further away.

5.4.2 Calibrated models

Although the difference between the parameter-equivalent models is a good benchmark, models are usually calibrated, which might increase or decrease the differences. Calibration obviously brings the traffic flows closer together in the calibrated point (in our case, where no links are tolled), but this could reduce the predictive power of the models.

Since we are only interested in the differences between models, and not in determining which mode best fits a particular dataset, we examine these effects by calibrating the NL and GNL models to the Probit route flows. Fig. 5.7 shows the flows on four representative links, resulting from the Probit, calibrated NL and calibrated GNL models.
5.4 Simulation results

As before, the differences between the models are small; in fact, the calibration has decreased them almost everywhere. It seems that, in this particular setting, a good choice of model parameters can compensate for the fundamental differences in the variance-covariance structure of the models. Moreover, the GNL model is now always closer to the Probit flows than the NL model is.

5.4.3 Overcalibration

The logit models used above are simple, and only have a few parameters that can be calibrated. Since, as we have seen above, calibration can bring the SUE models closer together, it might be advantageous to use a more flexible logit model with alternative-specific constants, as defined in Eq. 5.4. Fig. 5.8 shows the same Probit and NL flows as Fig. 5.7, but in addition, the flows resulting from a calibrated NL model with alternative-specific constants.

Naturally, the introduction of more calibration parameters allows the NL-ASC model to be closer to the Probit model when tolls are zero. If we move away from this calibrated state, however, the NL-ASC model is significantly further away from the Probit model than any other model we have examined, at least for the tolled link. This happens because the introduction of ASCs also changes the other calibrated parameters, $\beta$ and $\mu$. A decrease in $\mu$ makes routes less similar, and thus poorer substitutes. This decreases the price elasticity of a demand for each route. Conversely, a decrease in $\beta$ leads to larger differences in route costs, which increases elasticities. The introduction of ASCs may change both parameters in each direction. In this case, both $\beta$ and $\mu$ are lower, but the effect of the latter parameter is larger, which leads to an decrease in the price elasticity of demand for link 58. Hence, better
5 Stochastic user equilibrium traffic assignment with price-sensitive demand

Calibration of the initial equilibrium is not always good for out-of-equilibrium predictions, especially if there is no theoretical foundation (e.g., a correction for the number of left-hand turns in a route) for the addition of more parameters.

5.4.4 Implications for tolling

Although the differences in link flows between the various SUE models are small, it is difficult to say whether these small differences are important without specifying where these flows are going to be used for. We therefore examine two situations: one in which link 58 is tolled (in both directions) by a private operator, which maximizes its profits, and one in which the toll on this link is set by a social planner, which maximizes social welfare. Other policies, such as capacity extensions, are likely to give similar results, as they also affect the cost of using specific links (although indirectly, through a higher or lower congestibility, rather than directly).

Fig. 5.9 gives the private operator’s profit, as a function of $f_{58} = f_{85}$, for three models: MNP, calibrated NL, and calibrated NL-ASC. Naturally, the differences between the models are only important if they lead to different optimal tolls, or affect the choice for tolling as such. The Probit and NL models result in optimal tolls that are very similar; the profit levels are also comparable. If the Probit model is the correct model, using an NL model to obtain tolls would only reduce profits by 0.3%. Using an NL model with alternative-specific constants, however, would result in a significantly higher toll, and a profit loss of 8.6%, a direct result of the fact that this model uses a lower $\beta$ and $\mu$, which lowers the demand elasticity on link 58.
5.4 Simulation results

Fig. 5.10 shows the result of a similar exercise, in which a social planner maximizes the social welfare improvement resulting from tolling, as defined in Eq. 5.18. This figure is very similar to Fig. 5.9; again, there is only a minimal difference in optimal tolls between the MNP and NL models. If the Probit model is correct, using an NL model to obtain optimal tolls results in only a 1.2% welfare loss. Using the NL-ASC model, however, results in a significantly different toll, and a welfare loss of 5.4% if. This implies that, although simple logit-based SUE models often give very similar results to more flexible probit models, overcalibration can have a significant negative effect on policy effectiveness.

Figure 5.9: Profits as a function of tolls

Figure 5.10: Changes in welfare as a function of tolls
5 Stochastic user equilibrium traffic assignment with price-sensitive demand

5.4.5 Sensitivity

In Figs. 5.9 and 5.10, optimal tolls are higher in the MNP model than in the NL model, and even higher than in the ASC model. This is, however, not systematic. If, instead of link 58, another link is tolled, the results are different, as illustrated in Fig. 5.11, where we toll the link between nodes 10 and 11. Although this figure shows profits only, welfare follows a very similar pattern.

Because this link is much less central, it is used by fewer markets. Hence, although tolling still has a local effect, it has a much smaller impact on the other links in the model. As a result, the differences between the models are negligible. Although, here, both logit models result in a higher toll than the MNP model, this difference is so minor that the use of even the toll from an NL-ASC model results in less than a 0.04% profit loss if the MNP model is correct.

Figs. 5.12–5.14 show link flows and welfare for the various SUE models, in a situation where the congestibility of the links is lower (specifically, $\chi_l = \text{vot} \cdot \frac{K_l}{S_l} \left( 1 + 0.15 \frac{v_l}{V_l} \right)^2 + f_l$, where the square replaces the fourth power of the original BPR function). As Fig. 5.12 shows, the differences between models are larger than in the base case parameterization. This confirms that it is congestion that brings the models close together. As before, the GNL flows are much closer to the MNP flows than the NL flows. Moreover, as Fig. 5.13 shows, NL-ASC flows are now even further away from flows in the other models; in particular, the price elasticity of demand is much higher on the tolled link. As a result, the socially optimal NL-ASC toll is much lower, as Fig. 5.14 shows. If NL-ASC tolls are used while the MNP tolls are correct, this results in a 78% welfare loss, while using NL tolls would only reduce welfare by 28%.

---

Note that this does not only decrease the congestibility of all links, but through that, also affects demand levels, and hence all demand elasticities.
5.4 Simulation results

Figure 5.12: Lower congestibility – flows

Figure 5.13: Lower congestibility – overcalibration
5 Stochastic user equilibrium traffic assignment with price-sensitive demand

5.5 Conclusions

We have shown that, in a small but representative congestible transport network, simple logit SUE models can give results that are very similar to more general probit models. This result stems mostly from the fact that transport networks are congestible, which implies that the systematic utilities that users derive from overlapping routes are correlated, even if the random utilities corresponding to the routes are not. Moreover, in networks, OD-level differences can potentially cancel each other out. The differences in link flows between models are not systematic, i.e., one model does not always result in higher profit-maximizing or socially optimal tolls than another; this depends on the characteristics of the network and the links on which particular policies are enacted.

Since logit models are much less computationally intensive, this indicates that they might, at least in heavily congested settings, be a better choice. As logit models need no simulation, they can lead to more accurate results and allow for studying of more and more complex policy instruments and games (e.g., tax competition, networks with multiple operators, etc.). However, we also find that models can be overcalibrated, especially when parameters are introduced that have no theoretical justification. This is, for instance, the case if alternative-specific constants are introduced in a logit model when there is no theoretical basis for their inclusion. In that case, significant differences between models can arise, and careful evaluation of the various possibilities is necessary.

We have focused on a few representative SUE models, and have disregarded others; in particular, we have not examined any link-based route choice models, (e.g. Baillon and Cominetti, 2008; Fosgerau et al., 2013). Although these models are very useful for estimation, their complexity makes them less suitable for many simulation application. They do, however, have advantages; in particular, they avoid path enumeration, and do not need restricted choice sets. Further research has to determine how large different these models are from existing logit and probit models in practical situations.

Naturally, our results were obtained for a very specific situation. We have chosen our network such as to maximize the possible differences between models, and have chosen realistic parameters. It is unlikely that other networks or parameterizations would give significantly
different results, and our sensitivity analyses confirm this. The network we have examined, though small, features a large number of overlapping routes between all OD-pairs; moreover, many of these routes are used in equilibrium. We therefore do not expect different results, with stronger contrasts, for larger networks.
6 Modeling the formation of transport networks and its regulation

6.1 Introduction

The previous chapters have focused on short- to medium-term planning decisions in transport markets: pricing, scheduling, and the regulation of prices and schedules. Using the results from chapter 5, this final technical chapter analyzes long-term investment decisions. It focuses especially on private investment in transport networks, and the regulation of investment.

Privately constructed and operated railways have a long history, and private investors are now again increasingly being looked upon to fund the necessary new infrastructure (e.g. Debande, 2002; Estache and Serebrisky, 2004). The interest in private road supply is also increasing (Verhoef, 2008). However, if private operators not only control road prices or fares, but also decide which transport links are built, this could result not only in sub-optimal road prices or fares, but also in a network that looks much different than the social optimum. Even if, in this setting, regulators can not directly control all relevant variables, such as investment levels and fares, there may still be some opportunities for regulation; in particular, regulators may be able to either control the network structure or the fare levels. This chapter examines how this situation can be modeled, and illustrates, using a numerical example, how various regulatory strategies can address the allocative inefficiencies resulting from private infrastructure investment.

Three strands of literature have examined similar problems. First, the literature on transport network design, as it has been developed in the engineering and operations research communities, considers what an optimal transport network is, in certain settings. These studies usually look at the problem from the perspective of a single planner, which maximizes profits, or minimizes social or private costs (see e.g. the seminal work by Magnanti and Wong, 1984). If a social planner cannot directly decide on investment levels or prices, and these variables are instead set by separate, profit-maximizing operators that compete with each other, the problem becomes much more difficult; the network is then not designed by a single entity, but emerges as a result of competition.

Second, the strategic formation of networks has been examined extensively in the context of social networks (see Jackson, 2004, 2010 for an overview). The models developed in that strand of literature are often very similar to transportation models, in that links are constructed strategically by several players, where the construction of one new link effects payoffs in the whole network. However, there is one important difference: in social networks, ‘node’ is usually synonymous with ‘player’, and hence, if links are constructed strategically, the nodes themselves decide which links are constructed. In a transportation setting, this corresponds to a situation in which transport users living in a certain place decide where to build new transportation links to. Although, in some settings, this may be appropriate (e.g.

---

1This chapter is based on joint work with Erik Verhoef and Vincent van den Berg. I thank Hugo Silva, and the participants of the 2013 NECTAR meeting for helpful comments on earlier versions.
if infrastructure investment is the responsibility of local governments), this is generally not realistic. Transportation operators instead form a separate modeling level: they maximize their own profits, and do not directly act in the interests of the users traveling between the nodes that a new link would connect.

A third strand of literature looks at infrastructure investment from a more theoretical economic perspective. Most of these studies use simple models, often consisting of only one or two links (e.g., De Borger and Van Dender, 2006). In these models, investors only choose how much capacity to construct, and how to price this capacity. Moreover, capacity is usually assumed to be a continuous, differentiable variable. In many of these studies, optimal pricing can induce a first-best welfare-optimizing outcome; in some cases, capacity subsidization is also needed. Even in the few studies that do look at networks (e.g., Silva et al., 2014), this still holds. In reality, the economics of transport infrastructure investment are more complex. Network effects play an important role: individual links are complements and substitutes for others, and cannot always be considered in isolation. In addition, capacity is rarely completely continuous: at the very least, there are minimum and maximum capacities, and sometimes (as for railways) capacities may take only a few discrete levels.

Combining parts of these three strands of literature, we propose a modeling methodology to analyze investment decisions in a network, where capacities are discrete and decided by profit-maximizing competitors. Specifically, we examine a situation in which investment choices are binary variables: either a link is built, and in that case it will have a constant, exogenous capacity, or it is not built at all. We examine several game-theoretical structures:

1. Unregulated equilibria with one or more private investors deciding which links to construct, and subsequently, which fares to charge for access.

2. Fare-regulated equilibria with one or more investors deciding which links to construct, followed by a second stage in which a social planner decides on fares. We distinguish between two versions of this game, one in which the operators anticipate the socially optimal fare, and one in which they naively expect to set their own profit-maximizing fare when deciding on investment. Analyzing these two versions separately will allow us to isolate the effect of second-stage behavior on first-stage choices. Moreover, in real-world markets, unanticipated regulation is a plausible option.

3. Network-regulated equilibria, in which a social planner decides which links will be constructed. This is followed by a second stage, in which one or more private operators set fares. As with fare regulation, we distinguish between two versions: one in which the social planner correctly anticipates competitive fares (such that it selects the second-best network), and one in which it naively expects socially optimal fares (such that it selects a quasi-first-network). Again, these two versions allow for a better analysis of the various effects at play, and also reflect the complexities of real-world regulation.

4. Socially optimal investment and fare-setting. Here, social welfare is maximized; this serves as a benchmark against which we compare the performance of the other games.

With information about the outcomes of these games, it is possible to say which form of second-best regulation (network or fare) is best in which situation, how efficient they are,
6.2 Modeling methodology

6.2.1 User costs and demand

The lowest level of aggregation in most transport models is formed by individual transport users. To evaluate the effects of pricing or regulation, it is important to correctly model how users decide whether they will travel, and if so, which route they will take. Hence, the choice of user cost functions is, naturally, key. There is a wealth of literature on these user cost functions, mostly for road travel, but also for other modes. One of the most popular cost functions for road travel is the BPR function (US Bureau of Public Roads, 1964), which assumes that travel times are a (non-linear) function of the traffic flow on a link, given its capacity. These traveling times can then be multiplied with users’ values of time (VOT) to obtain user costs.

In a public transportation network, travel times (other than, possibly, boarding times) do not generally depend on usage levels. Instead users experience different forms of congestion during their trip, including in-vehicle and on-platform crowding. There are many empirical studies that attempt to estimate the costs of on or both these forms of congestion; for an overview, see Wardman and Whelan (2011) and Li and Hensher (2011). A popular specification, proposed by Whelan and Crockett (2009), uses a crowding multiplier, which multiplies the value of in-vehicle time, and hence, ignores the relatively unimportant on-platform crowding. As their analysis shows, a linear specification fits the data well once the occupancy rate, in terms of seats, exceeds 100% (such that all additional passengers stand). As long as there is still a significant number of seats left, crowding costs increase at a much lower rate, but they are still close to linear. Hence, as long as it is reasonable to assume that a negligible number of passengers can sit down or a negligible number of passengers stands (as in trains where a reservation is required), using a single linear cost function seems reasonable. Given that, the generalized price of using link \( l \), \( \chi_l \) is then:

\[
\chi_l = VOT \left( 1 + Av_l \right) T_l + f_l
\]

where \( VOT \) is the value of time, \( v_l \) the number of users taking link \( l \), \( f_l \) the fare, \( T_l \) the travel time, and \( A \) a parameter. This parameter can be used to calibrate the model, such that relative to the first-best socially optimal situation, how the networks differ between the games, etc.

In the next two sections, we explain the building blocks of a model that can be used to analyze these games, and discuss the economic theory behind the potential outcomes. To illustrate the challenges and possibilities of this methodology, we construct a numerical model in section 6.4, using a simple network with a limited number of nodes and potential links, and use this to simulate the four games listed above. In the numerical simulation, we fix the number of links, to specifically examine the allocative efficiencies of the games. From the results of this simple simulation exercise, we can already draw some economic conclusions that are likely to hold in real-world applications. Most importantly, second-best fare regulation is not always enough to achieve welfare-maximizing outcomes, even if the regulator can also set the total amount of capacity that operators build (through, for instance, location-independent per-unit capacity subsidization – this subsidy can be set such that the optimal amount of total capacity is attained, but it is unlikely that this capacity is constructed in the correct location). Section 6.5 concludes.
the flows correspond to actual flows, or such that the crowding multipliers \((1 + A \cdot v_l)\) have realistic values. The empirical studies mentioned above report estimates that are roughly between 1 and 4, depending on mode-specific details. It is natural to assume that the cost of taking a given route \(r\) is simply the sum of the the link-based costs over all links that make up the route:

\[
c_r = \sum_l a_{rl} \chi_l
\]

(6.2)

where \(a_{rl} = 1\) if link \(l\) is part of route \(r\), and zero otherwise.

In many cases, the total number of travelers between two nodes will depend on the cost of travel, and thus, demand is price-sensitive. There are several ways to model this sensitivity to prices. One way is to express the total travel demand between two nodes as a direct function of the (expected) cost of travel. Another is to keep the total amount of users fixed, but to include the option not to travel as a discrete choice in the route choice model. The latter option is often easier to include in a model, as the next section will show. It may be difficult to determine what the (exogenous) value of this no-travel alternative should be; however, finding the parameters of an explicit demand function will also be difficult. Since the cost of the outside alternative directly affects the flows on all links, it can be treated as a calibration parameter. Following the previous chapter, the user equilibrium model below, and our numerical simulations, will use this no-travel alternative approach, rather than an explicit demand function that depends on the expected route costs.

### 6.2.2 User equilibrium

Having specified route-based cost functions \(c_r\), the next building block is a traffic assignment model, which calculates the usage levels of all routes in equilibrium for any set of fares \(f_l\).

Here, several approaches are possible. A deterministic user equilibrium can be derived from Wardrop’s principles (Wardrop, 1952), which specify that all used routes must have the same generalized prices, which is lower than the cost of all unused routes (i.e. for all routes \(r\) and \(s\), \(0 \leq c_r - c_s \perp v_r \geq 0\)). Hence, the solution to these conditions gives a Nash equilibrium if the number of users is large enough: no single users can decrease its costs by unilaterally choosing a different route.

The Wardropian user equilibrium has been proven to exist, and to be unique, for a very general class of cost functions; it can be approximated using, for instance, a Frank-Wolfe algorithm (Dafermos and Sparrow, 1969). In some situations, however, its properties are unattractive. It may result in a large number of unused routes through a network, even if usage data suggests these routes are used; this disparity is difficult to address through calibration of the model parameters. Moreover, the usage levels \(v_r\) are not continuous in user costs: for every route, there is a point at which costs become so high that suddenly, the route will cease to be used. This is particularly problematic for the analysis of competition in networks, because it implies that the exploitation profits of a given link are also not continuous in route costs. If there is more than one operator, and each operator sets a fare on its own link, Nash equilibria in fares may be difficult to calculate, or may not even exist.

Partly as an answer to these problems, stochastic user equilibrium (SUE) assignment models have become popular. In those models, user costs consist of the deterministic costs \(c_r\), and a stochastic term. This stochastic term is not observed, but is assumed to have a known distribution. It is then possible to calculate the probability that a randomly selected user will take a given route; summing over all users will give the expected total usage level. An
attractive property of SUE models is that usage levels will never reach zero, but asymptote towards it; moreover, they can easily be calibrated to real-world data.

Many different SUE models have been developed, each with its own assumptions about the stochastic part of user costs; several of these have been examined in chapter 5. The simplest version, the multinomial logit (MNL) model, assumes that the stochastic term of each route is independent of all others, and follows a Gumbel distribution. The route choice probabilities are then easy to calculate, and have a closed-form solution. At the other end of the scale, a multinomial probit model allows for any type of correlation between the error terms of all pairs of alternatives; this model can only be solved with numerical methods.

In the context of a larger model that analyzes competition, an SUE model that has closed-form probabilities is preferable, as it allows for much shorter computation times. However, in a network setting, routes usually partially overlap. This implies that the deterministic route costs correlate; it is then unrealistic to assume that the stochastic terms are uncorrelated. As the previous chapter has shown, generalized nested logit (GNL) models offer a useful compromise between realism and computational intensity. In this model, the number of nests is equal to the number of links plus one (for the no-travel alternative); each route may belong to a number of nests, where the shares of each route attributed to nests are calculated using free-flow travel times. Route choice probabilities are then given by Eq. 5.7.

If the physical routes only differ in length, and not in other characteristics, it is reasonable to set all $\mu_l = \mu$. Given Eq. 5.7, the expected total travel flow on each link is

$$v_l = \sum_{ij} \sum_r a_{ijr} a_{ir} D_{ij} p_{ijr}$$

(6.3)

where $D_{ij}$ is the potential demand, or demand intercept.

### 6.2.3 Operator profits and social welfare

Operator income is the product of fares and link-based demand, summed over all the links that the operator has constructed. In general, operators also have operating and maintenance costs, which may depend on the usage levels of the links. In our numerical simulations, we will disregard these costs, as they unnecessarily complicate the analysis. Naturally, the total link flow $v_l$ is a function of all route probabilities $p_{ijr}$, which themselves depend on the flows through congestion. It is theoretically possible to substitute the route probabilities in the profit functions, and to solve to find a fixed point. The resulting profit functions will then be a function of $f_l$ only, and can be maximized. Since the route choice probabilities follow a relatively complex exponential function, it is more convenient to treat all $v_l$ as decision variables, such that firm $x$ maximizes its profit $\pi_x$:

$$\pi_x = \sum_{l \in L_x} (f_l v_l - c_l)$$

(6.4)

where $c_l$ is the investment cost associated with the construction of link $l$, and to then impose Eq. 6.3 as a constraint. Since each firm solves this problem, this will give the same Nash equilibrium fares.

A social planner instead maximizes social welfare which, in this case, is the sum of the consumer surplus and the sum of all operators’ profits. In a discrete choice model, such as the GNL model discussed above, this consumer surplus can be calculated by using the
6 Modeling the formation of transport networks and its regulation

logsum, or expected utility. In the GNL, this logsum is given by

\[ \frac{1}{\beta} \log \left( \sum_l \left( \sum_{r'} a_{ijr'} \left( \alpha_{r'd} e^{-\beta c_{r'}} \right)^{\frac{1}{\mu_l}} \right)^{\mu_l} + e^{-\beta C} \right) \]  

(6.5)

for a user traveling between nodes i and j. If no links are constructed, this simplifies to \( \log (e^{-\beta C}) = -\beta C \). Consumer surplus for the total number of users traveling from i to j can, therefore, be calculated as the difference in logsums between a situation in which no links are built, and the situation of interest\(^3\), multiplied by the number of individual decision makers \( D_{ij} \) for that market. Summing these values over all markets gives the total consumer surplus. Given our assumptions about operator profits, total welfare \( W \) is then given by:

\[ W = \sum_{ij} D_{ij} \left( \frac{1}{\beta} \log \left( \sum_l \left( \sum_{r'} a_{ijr'} \left( \alpha_{r'd} e^{-\beta c_{r'}} \right)^{\frac{1}{\mu_l}} \right)^{\mu_l} + e^{-\beta C} \right) + \beta C \right) + \sum_x \sum_{l \in L_x} (f_l v_l - c_l) \]  

(6.6)

6.2.4 Optimal investment, fares, and regulation

The building blocks explained above can, together, be used to compute usage levels, profits, and welfare, given a set of fares, and a set of links. This section discussed how operators and regulators set these variables. We will do this for the four situations mentioned above: an unregulated setting, where one or more private operators decide which links to construct, and which fares to set; a setting with network regulation but private profit-maximizing fares; a setting with a private profit-maximizing network but regulated fares; and a fully regulated setting, where a social planner decides on fares and investment. In the next section, we will discuss how the resulting equilibria can be obtained with numerical methods.

No regulation

This is a two-stage game, which can be solved by backwards induction. In the second (pricing) stage, each operator solves

\[ \max_{f_l, v_l | l \in L_x} \pi_x = \sum_{l \in L_x} (f_l v_l - I_l) \]  

(6.7)

s.t. Eq. 6.3 \( \forall l \), where \( x = 1 \) if there is one monopolistic operator, and \( x \in \{1, \ldots, X - 1, X\} \) if there are more, and \( I_l \) is the construction cost of link \( l \). In our numerical simulations, we will assume that, if there is more than one operator, \( X \) is exogenous. In general, if entry is allowed, \( X \) should reflect the maximum number of potential entrants in the market, or the maximum number of links, whichever is lowest. Each operator independently determines which fares to charge for the use of its own infrastructure; naturally, since \( v_l \) is a function of all fares, its profits are influenced by its competitor’s decision. In the first (investment) stage, each operator chooses a set of links \( L_x \) to maximize its profits.

\[ \max_{L_x} \pi_x (f_l^*, v_l^*) \]  

(6.8)

\(^3\)The logsum is an indicator of users’ expected utility (including utility derived from choosing the no-travel option; hence, to calculate the benefits of transportation, the expected utility in a setting where all users decide not to travel is subtracted.
6.2 Modeling methodology

\[ L_x \subseteq \left( L \setminus \bigcup_{x \neq x} L_x \right) \tag{6.9} \]

where \( \{ f^*_x, v^*_x \} \) is the solution to the second-stage problem. Again, \( x = 1 \) if there is one monopolistic operator, and \( x \in \{ 1, ..., X - 1, X \} \) if there are multiple operators. Constraint 6.9 ensures that each link can only be constructed once. There may be other restrictions on \( L_x \); each operator may, for instance, only be allowed to construct one link (i.e., \( |L_x| \leq 1 \)), as we will assume in the numerical simulations below.

If there are multiple operators, it is natural to assume that all fares are set simultaneously, and that \( \{ f^*_x, v^*_x \} \) is the pure-strategy Nash equilibrium; no operator has an incentive to unilaterally change its fare.\(^4\) Similarly, in the first stage, we consider all equilibria in which no operator has an incentive to unilaterally choose another set of links. Contrary to the user equilibrium, the network investment equilibrium may not always exist even if cost functions are highly simplified. There may also be more than one Nash equilibrium.

### Price regulation

Again, this is a two-stage game, which can be solved by backwards induction. Now, in the second stage, the social planner solves

\[
\max_{f_l, v_l} W \tag{6.10}
\]

s.t. Eq. 6.3 \( \forall l \). The first (investment) stage problem now depends on whether the operators anticipated socially optimal prices, or expected to set their own profit-maximizing prices. In the latter case, they solve the same first-stage maximization problem as they would in the unregulated market. If socially optimal prices are anticipated, they solve a similar problem, but with \( \{ f^*_x, v^*_x \} = \arg \max_{f_l, v_l} W \). In either case, as before, there may be one pure-strategy Nash equilibrium, or none, or several.

### Network regulation

Here, as in the unregulated equilibrium, operators maximize their profits s.t. Eq. 6.3 \( \forall l \) in the second stage. The first stage, in which the social planner designs the network, depends on whether it expects to also set prices. If both cases, it solves

\[
\max_{L_x} W \tag{6.11}
\]

s.t. \( L_f \subseteq L \). When it incorrectly assumes socially optimal pricing will follow (a quasi-first-best situation), it uses \( \{ f^*_l, v^*_l \} = \arg \max_{f_l, v_l} W \). If, instead, it selects the second-best network, \( \{ f^*_l, v^*_l \} = \arg \max_{f_l, v_l} \pi_x \)

\(^4\)In our context, mixed-strategy Nash equilibria are difficult to justify. Other equilibrium concepts may be more realistic alternatives; e.g., a Strong Nash equilibrium, which is also robust against deviations of more than one player or a Coalition-proof Nash equilibrium, which is robust against deviations of more than one player where these players cannot make binding commitments. In most of the transportation literature, however, simple Nash equilibria remain the most popular way to model simultaneous decisions.
6 Modeling the formation of transport networks and its regulation

6.2.5 Solution methods: mixed-integer programming

The games outlined above can be formulated as mixed-integer problems. These types of problems are very common in other markets, and consequently, there is a large literature on solution methods and algorithms.\footnote{E.g., unit commitment problems in electricity markets, where generators decide which generators to turn on, and participate in a common market afterwards (Hobbs et al., 2001).} The difficulty of transport network formation is that is that each potential link can be built by each potential operator; there is not necessarily any ownership before investment.

One way to handle this is to define a variable $I_{xl}$, which is equal to one if operator $x$ owns link $l$, and zero otherwise.

$$I_{xl} \in \{0, 1\}$$  
(6.12)

$$0 \leq \sum_x I_{xl} \leq 1$$  
(6.13)

Moreover, if each operator is only allowed to construct a maximum of one link,

$$0 \leq \sum_l I_{xl} \leq 1$$  
(6.14)

Further, define the fares and usage levels as

$$f_l = \sum_x F_{xl}$$  
(6.15)

$$v_l = \sum_x V_{xl}$$  
(6.16)

where $F_{xl}$ and $V_{xl}$ are variables and

$$F_{xl} (1 - I_{xl}) = 0$$  
(6.17)

$$V_{xl} (1 - I_{xl}) = 0$$  
(6.18)

Naturally, flows also have to be consistent with the user equilibrium, as given by Eq. 6.3.

If there is only one operator, or a social planner decides on all variables, the problem is then a mixed-integer nonlinear program (MINLP); the operator or social planner chooses investment levels $I$, fares $F$, and flows $V$, subject to the user equilibrium constraints, and the constraints listed above. If there are more operators, it is an equilibrium problem with equilibrium constraints (EPEC), which is generally very difficult, though not always impossible to solve. Using the new variables $I$, $F$ and $V$, the second (pricing) stage can be replaced by its Karush-Kuhn-Tucker (KKT) conditions; these can then be used as constraints in the first stage. Since the KKT conditions represent an equilibrium (i.e., for a given investment level, there is only one optimal second-stage solution), the feasible region of the first-stage problem is necessarily non-convex, even without the integer constraints. This makes solving the first-stage problem very difficult. Moreover, the user equilibrium constraints involve terms that are summed over all routes, and the set of routes that serve each market changes with the network configuration. Formulating the user equilibrium constraints is, therefore, far from easy.

In many cases, however, the number of potential network configurations is limited. Moreover, it may be possible to further reduce this number by discarding configurations that are...
highly unlikely to be equilibria (e.g., configurations where one link connects two nodes between which travel demand is very low, and there are no other links connecting these nodes to larger markets). In that case, it may be possible to solve the second-stage for each possible configuration; the first-stage problem then reduces to choosing which equilibria lead to the highest social welfare, or which equilibria are stable to unilateral deviations in investment decisions. Hence, instead of replacing the second stage of the problem by its KKT conditions, the first stage is replaced by an exhaustive search. We will further explore this method in the next section.

6.2.6 Solution methods: exhaustive search

If a social planner or a single private operator chooses which links to construct, an exhaustive search over the whole set of potential equilibrium networks is a straightforward way to eliminate the investment stage, and with it, all integer variables. As explained above, this set of potential solutions may not be very large, even if the set of all possible network configurations is much larger; many network configurations can be discarded ex-ante. Once the second (pricing) stage has been solved for all potential network equilibria, choosing the configuration with the highest profit or social welfare is simple.

If multiple operators invest, an exhaustive search is more complex. In this case simply solving for all potential equilibria is not enough. Whereas, in the case where one operator or social planner invests, the network selection simply amounts to choosing the configuration with the highest profit or social welfare, the selection procedure is more complicated here, as each operator now has its own profit. If, as we have assumed above, the operators compete in a Nash setting, a network configuration is an equilibrium if and only if no single operator has an incentive to change its investment strategy. Hence, to examine whether a network configuration is indeed an equilibrium it is not only necessary to solve the pricing stage for that configuration, but also for all configurations in which one of the operators chooses a different link or set of links. If at least one of these unilateral deviations is profitable for the operator that deviates, the original allocation is not a Nash equilibrium. In most cases, however, it is not necessary to solve the pricing stage for all possible deviations; as soon as one profitable deviation is found, it is not necessary to look further. Hence, in a situation where the investment stage is competitive, the exhaustive search can be performed as follows:

1. Define a set of potential network equilibria $S$, containing all allocations $L$ that might be Nash equilibria. Here, each allocation describes which operators construct which links. The set of links constructed by an operator may be empty; naturally, this implies that its profits are zero.

2. Solve the second (pricing) stage for all allocations in $S$. This step can easily be parallelized.

3. Check the allocations in $S$ against each other; if any allocation represents a unilateral deviation from another, and this deviation is profitable for the operator that deviates, this other allocation is not a Nash equilibrium, and can be discarded.

4. Define the set of equilibria that have not been discarded as $S^1 \subseteq S$.

5. For each $L$ in $S^1$, sequentially solve the second stage for all allocations that represent a unilateral deviation from $L$. If a deviation is profitable, discard that $L$, and move to the next. Note that leaving the market and not constructing any links may also be
6 Modeling the formation of transport networks and its regulation

a profitable deviation for an individual operator; the same is true for the decision to construct a link by an operator that, in the original $L$, did not construct any links. These last two types of deviations therefore also need to be checked.

6. If one or more $L$ are left, they are Nash equilibria.

Although the pricing stage will have a unique equilibrium for any reasonable set of congestion cost functions, the investment stage could have a single equilibrium, several equilibria, or no equilibrium at all. In our numerical simulations, we always find a single unique equilibrium, but this is at least partially because we limit the number of operators. In a large network, with many operators, it is more likely that multiple network equilibria exist.

6.3 Theory

Before we illustrate the uses of the above models in a numerical example, it is useful to briefly consider which results might be expected, based on economic theory. Two types of results are of interest here: the performance of monopolistic and competitive markets, relative to the social optimum, and the efficiency of the various regulatory strategies.

Naturally, a monopolistic market, in which one operator invests in new links and subsequently sets fares, will not achieve the maximum social welfare. If the capacities of new links were continuous variables, a monopolist, using the same investment rule as a social planner but setting higher fares, would always invest less than optimal (see Small and Verhoef, 2007); this happens because the monopolist, ignoring any consumer surplus, evaluates the investment rule at a lower usage level. The monopolistic markup decreases with the price elasticity of demand. In our user equilibrium models, this elasticity is a function of the cost of the no-travel alternative, the marginal utility of money, and the nesting coefficient. If the no-travel alternative is more costly, demand for travel is less sensitive to fare levels; the same is true if the marginal utility of money is lower. Moreover, if overlapping routes are perceived as very similar alternatives, users are more likely to choose the no-travel alternative, relative to a situation in which the routes are unrelated.

In our models, capacities are discrete. Compared to the continuous case, this may increase or decrease social welfare. If, for a given link, the monopolist’s optimal capacity choice in a setting where capacity was a continuous variable is lower than the capacity that can be chosen in a discrete setting, it can choose not to construct the link, which will decrease social welfare, or it can choose to construct it at the higher capacity, which will increase social welfare.

A competitive market, in which several operators construct links and set fares, may lead to a higher social welfare than a monopolistic market. However, depending on the network structure, it could also lead to a lower social welfare. In the pricing stage of the models, the existence of parallel links will decrease fares and increase social welfare, since parallel links are direct substitutes for each other, and this substitutability increases competition. If competing operators own serial links, which are complementary to each other (i.e., some travelers use two or more to get to their destination), this increases fares relative to the monopoly, as operators disregard the negative effects of a fare increase on other operators’ profits (Economides and Salop, 1992). This, in turn, decreases welfare. Naturally, in a larger network, links can be substitutes for some users, and complements for others; in that case, the net effect depends on the relative sizes of the markets.
In our settings, however, the network structure is endogenous. Whereas, in a monopolistic market, the degree of parallelity will generally be low (as the construction of a link parallel to another will negatively effect profits on the original link), it may be much higher on a competitive market (as competitors disregard this negative effect). Conversely, the degree of complementarity will be lower with competition: competitors disregard the positive effects of the construction of a new link on the profits made on other links that connect to it. Moreover, the threat of entry by other operators generally increases total investment levels (Schmalensee, 1978; Scotchmer, 1985; van der Veer, 2002). All this suggests that the positive effects of competition may, in equilibrium, outweighs the negative effects. Whether this is true or not for a given network depends on the demand parameters mentioned above, and on the set of possible links: if it is only possible to construct links that are complements for most users, for instance, competition may not increase welfare. Moreover, since capacities are decided on before fares are determined, competitors may be able to avoid competition through underinvestment, because high levels of congestion make fares less important (de Palma and Leruth, 1989). As in the monopolistic case, the discreteness of link capacities may make things worse or better, depending on the available capacities and their costs.

Naturally, if regulators can set fares and decide which links will be constructed, the first-best social optimum can be achieved. The same is true if regulators set fares and subsidize investment by posting a separate subsidy for every link. If first-best regulation is not possible, there are three types of distortions: the total amount of capacity may be suboptimal, the capacity may not optimally located in the network, and fares may be too high. Capacity regulation would fix the first two; fare regulation would directly address the third, but indirectly also affect the network formation. A blanket capacity subsidy, independent of location, could optimize the total capacity, but not its location, nor fares that are charged for access.

Whether network regulation or price regulation is better depends on the model parameters. Network regulation has one important advantage in competitive markets: it eliminates the probability that a highly undesirable Nash equilibrium network will emerge. Since, as we have seen, it is possible that multiple networks are Nash equilibria, operators may get stuck in an equilibrium that is very far away from the optimum. Fare regulation cannot address this directly. If the distortion in total capacity investment is small, fare regulation is likely to be more efficient. When operators decide which links to construct, their objectives are not perfectly aligned with those of the social planner, but they are similar. Operators want to invest in places where they can make the highest profits, while social planners would invest in places where the highest social surplus gains can be achieved; both of these conditions are likely to hold in the largest markets. When operators decide on fares, however, there is a much larger difference in objectives. By definition, quasi-first best network regulation, which ignores the fact that fares will be set competitively, can never result in a higher social welfare than second-best network regulation. Similarly, non-anticipated second-best price regulation can never be worse, and is likely to perform much better, than anticipated regulation.

The discrete nature of investment does make a difference for regulation. It has been shown that, in many cases, optimal fares can often restore first-best outcomes if capacities are continuous (see e.g Silva et al., 2014): if fare regulation is not enough, the addition a location-independent capacity subsidy will usually fix everything. If investment is discrete, this is not longer the case, as allocative inefficiencies may still remain. Given the partial alignment of private operators’ and social planners’ objectives, however, these inefficiencies are likely to be small; we will further explore them in the numerical simulations below.
6 Modeling the formation of transport networks and its regulation

6.4 Numerical simulations

To illustrate how the model proposed in the above sections behaves, and which economic conclusions can be drawn from it, this section will present the results of a numerical simulation, using a small but representative network. To limit the computational intensity of the models, we only consider the allocative inefficiencies of the various market structures. To that end, we assume that construction costs are the same for all links, and that the total number of links is fixed. We then examine where these links are constructed, and which fares are charged for access, in each of the four situations mentioned above (no regulation, price regulation, network regulation, full regulation), with myopic and forward-looking agents, in a monopoly and an oligopoly where each operator only owns one link. The fixed number of links can represent a situation in which the regulator has decided on the optimal total capacity level, and subsidizes capacity at a rate which induces operators to invest at this optimal capacity, without spatially differentiating the subsidies.\(^6\) We use the exhaustive search approach outlined above, solving the second stage of the models using the KNITRO nonlinear solver (Byrd et al., 2006) in Wolfram Mathematica 8.

6.4.1 Parameterization

We test our models on the simple eight-node network shown in figure 6.1. In this figure, lines represent possible connections, and the sizes of the nodes indicate their actual sizes in our base case parameterization. Although this network is relatively small, it still features a large number of possible routes, many of which partially overlap. These routes not only exist in theory, but are also used by a significant number of passengers in the equilibria we examine. A smaller network would not be suitable for our purposes; the number of links and routes would be too limited to properly examine competition between operators. The middle node is purposely slightly offset, to avoid situations in which the equilibria are not unique only because links have exactly the same expected costs and benefits; the average link length is 40km.

![Figure 6.1: Network](image)

The realized demand for travel on any constructed link will, of course, depend on prices,

\(^6\)In general, the optimal number of links may depend on the market structure; our setup ignores this, and fixes the number of links across all models.
6.4 Numerical simulations

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_i$</td>
<td>150000 0 120000 225000 200000 375000</td>
</tr>
<tr>
<td>$A$</td>
<td>1/8000</td>
</tr>
<tr>
<td>$B$</td>
<td>1/50000</td>
</tr>
<tr>
<td>$C$</td>
<td>40 $</td>
</tr>
<tr>
<td>$VOT$</td>
<td>10 $/h</td>
</tr>
<tr>
<td>$V$</td>
<td>120 km/h</td>
</tr>
<tr>
<td>$\beta$</td>
<td>1</td>
</tr>
<tr>
<td>$\mu$</td>
<td>1/2</td>
</tr>
</tbody>
</table>

Table 6.1: Base case parameters

and the levels of congestion, as shown below. We use a simple gravity model to calculate the ‘potential’ demand, or demand intercept $D_{ij}$, for travel between nodes $i$ and $j$, such that:

$$D_{ij} = B \frac{S_i S_j}{L_{ij}}$$ (6.19)

such that the potential demand between two nodes increases with each node’s size $S_n$, multiplied by a parameter $A$ and decreases with the Euclidean distance $L_{ij}$ between the two nodes. Travel costs are assumed to increase linearly in the usage levels, and route choices and actual numbers of travelers are modeled with a GNL SUE. Travel times $T_l$ are calculated by dividing the link length by the travel speed $V$. All parameters are summarized in table 6.1.

6.4.2 Monopoly

We first consider a situation in which there is only one operator ($X = 1$), who chooses which four links to construct out of all possible links shown in Fig. 6.1. As explained above, this implies that the results will only show the effects of potential suboptimal capacity locations, not of a possible suboptimal total capacity. Starting with the base case equilibrium, we increase the size of the central node, node 5, examining in particular the social welfare in the various regulated and unregulated monopolistic equilibria, relative the social welfare in the first-best equilibrium where the both the set of links and the fares are socially optimal. Fig. 6.2 shows the results of this exercise; the vertical axis measures social welfare relative to welfare in the first-best situation for the five cases: an unregulated monopoly (Private), anticipated and unanticipated second-best fare regulation (SB f-A and SB f), and quasi-first-best and second-best network regulation (QFB net and SB net). We plot this relative welfare for six different values of $S_5$.

Several interesting observations can be made here. First of all, for a wide range of $S_5$, second-best fares are not enough to achieve the first-best solution. Since the relative welfare is smaller than one, there is still some allocative inefficiency: the network configuration is suboptimal. Since we assume that the total capacity equals the optimal number of links this means that even capacity subsidization and second-best pricing together are not sufficient.

---

7See the appendix for an overview of the crowding multipliers in various simulations.
Only if capacity subsidies are different for each potential link can the first-best welfare be attained. This is due to the discreteness of investment. Fig. 6.3 shows the two networks that are used in the various equilibria. If node 5 is very small \( S_5 \leq 100000 \), both a private operator and the social planner would choose network (a). If we increase the size of this central node, it becomes socially optimal to instead construct network (b), connecting node 3. The private operator, however, maximizing only its own profits, continues to prefer network (a) in the whole range of \( S_5 \) examined in Fig. 6.2. Only if \( S_5 \) were to increase even further would the private operator’s network coincide with the social optimum.

However, although the first-best situation can not always be achieved, the welfare losses are small: the private operator has no incentives to construct a network that is very far away from the optimum. Even in the worst case, with \( S_5 = 20000 \), the welfare in an unregulated situation is only a few percent lower than the first-best welfare. Note that the number of links is fixed here: there are no inefficiencies in the total amount of investment. If these would be taken into account, the monopolistic outcomes would probably be much further away from the optimum.

Secondly, quasi-first-best network regulation can decrease welfare, relative to an unregulated situation. This has important policy implications: network regulation can only be successful if the regulator knows exactly how fares will be established after the network has been constructed.

Thirdly, in this particular example, second-best fare regulation always results in a higher social welfare than second-best network regulation. Although, for most of the parameter space, the private network is not optimal, this can to a large extent be corrected with second-best fares; the other way around, it is more difficult to pre-empt suboptimal fares with the limited amount of options for second-best network investment that are available. Naturally, this result depends on the model parameters, and as we will see below, the opposite might happen. Anticipation of fare regulation has no impact here, because of the highly discrete nature of capacity investment, and because, given the price-regulation, the monopolist always chooses the optimal network in this parameterization.

To see how these results depend on the model parameters, we fix the size of the central node to 200000 and vary the value of time from its initial value of 10$/hr; Fig. 6.4 shows the results. As the value of time increases, the monopolist sets higher fares, and still chooses a network that is sub-optimal; the relative performance of an unregulated monopoly decreases. Interestingly, however, as the value of time increases, network regulation also becomes more attractive, and for a value of 11$/hr, second-best network regulation is even better than second-best fare regulation. This implies that, in the second-best optimal network, private operators set fares that are closer to the optimum as values of time increase, even though, in the private network, private and optimal fares diverge further. Naturally, this result is specific to this particular model, but it does show that one second-best regulatory strategy does not always dominate the other, and that these results are parameter-sensitive.

6.4.3 Competition

Fig. 6.5 shows the results of a similar exercise as Fig. 6.2, but this time for a situation in which there are four competing operators \( X = 4 \). Each operator chooses which link to construct, out of all the possible links in Fig. 6.1 The results are very similar; in most cases, allocative inefficiencies are still present, even if second-best fares are implemented. The network choices are different: in an unregulated Nash equilibrium, network (b) in Fig. 6.3 emerges if node is 5 smaller than 250000, while network (a) emerges for larger \( S_5 \).
6.4 Numerical simulations

Figure 6.2: Monopoly – varying $S_5$

Figure 6.3: Monopoly – networks
results in situations where welfare in an unrestricted competitive market is lower than in an unrestricted monopoly. This has everything to do with the network structure: given the small network, and the small number of links, there are few opportunities to build parallel links. Hence, the networks that emerge are exclusively serial. As discussed above, serial competition increases fares, as operators disregard the negative effects of a fare increase on their competitor’s profits; this decreases welfare, relative to the monopoly. Although this does not directly affect the games in which fares are set by the social planner (SB f and SB f−A), it does reduce welfare in the games where operators set fares.

In this competitive case, anticipation does play a role: first-best outcomes can sometimes be achieved if operators expect to set competitive fares, but are faced with regulated fares after they have chosen their links. The range in which this happens is relatively small, however; if $S_5$ is very large or very small, the fare-setting mechanism does not influence first-stage decisions. Quasi-first-best network regulation can still decrease welfare, relative to the unregulated case. The results are highly parameter-sensitive, more so than in the monopolistic case: although the first-best can be achieved if $S_5 = 150000$, this is not the case if the size of node 5 is decreased to 100000 or increased to 200000.

Fig. 6.6 illustrates the parameter sensitivity of the model further; in this figure, we again fix $S_5$ to 20000 and vary the value of time. Whereas, in the monopolistic case, a small change in the value of time only had a small impact on the relative performance of the various games, the sensitivity in this competitive case is twice as large. An increase in the value of time increases the relative welfare of an unregulated equilibrium because, in this particular case, it decreases the difference between private and socially optimal fares. It affects the performance of anticipated fare regulation in a non-monotonic fashion. As in Fig. 6.5, there are parameter values for which the network that emerges is the same regardless of whether fares are set by the regulator or the operators, but there are also parameter values for which
competitive fares and flows differ from second-best optimal fares and flows in such a way that operators choose another network.

The first data point in Fig. 6.6 also further illustrates the difference between second-best and quasi-first best network regulation. Even though, if \( VOT = 9 \), the private network that emerges is the same as the first-best socially optimal network (as unanticipated fare regulation achieves the first-best outcome), the regulator can achieve a higher social welfare by choosing another network, which will result in fares that are closer to the optimum. Hence, the second-best network is different from the private network, even if that private network would be optimal in a first-best setting.

### 6.5 Conclusions

In this chapter, we have proposed a two-stage methodology that captures the strategic formation of transport networks by one or more private operators. Specifically, we have examined a situation in which investment capacities are discrete: links can be constructed or not, rather than constructed with any chosen capacity. The models we have developed can be used to analyze settings where transport networks are not designed by a single actor, but emerge as a result of competition. We have examined several settings: an unrestricted monopolistic or competitive market, a market where only fares or networks are regulated, and a fully regulated setting. Underlying the investment and fare setting stages is a user equilibrium model, which determines how many users travel, and which routes they will take through the network. These models are difficult to solve, but, in some cases, integer programming or a smart exhaustive search can go a long way.

We have illustrated this methodology with a simple example, in which we have analyzed the
allocative inefficiencies of private transport infrastructure investment in a small network. As we have shown, these allocative inefficiencies are relatively small in a monopoly: although the private operator’s objectives are not fully aligned with social planner’s, the operator has no incentive to construct a network that is very different from the optimum. In our simple model, the same holds in a competitive market, where multiple operators own links; this, however, does not generalize to larger networks, where multiple equilibrium network configurations may exist, some of which are likely to be much further away from the optimum configuration.

The numerical simulations also show that, even if the total number of links is fixed to its optimum, and fares are set by the social planner, first-best outcomes cannot always be achieved. Because of the discrete nature if infrastructure investment, some allocative inefficiencies remain; these can only be addressed with direct network regulation. If second-best fare regulation is not anticipated, the parameter space in which the first-best optimum can be achieved without network regulation increases. If network- and fare regulation cannot both be implemented, the model parameters determine which second-best regulatory strategy is best.

Naturally, this is only a start. Although our simple model can already be used to draw some general economic conclusions, larger, more sophisticated models are needed to fully understand the impact of private investment in transport networks. The methodology we have discussed will, hopefully, be useful for this.

Appendix

Fig. 6.7 shows the distribution of crowding multipliers, over all routes and all examined network configurations, for the different sizes of node 5 that have been examined.
6.5 Conclusions

Figure 6.7: Distribution of crowding multipliers
7 Conclusions

7.1 Results and implications

In general, two different types of conclusions can be drawn from the results presented in economic theses like this one. First, there are practical conclusions, which can inform policy. Second, there are theoretical conclusions, which confirm or disprove existing theories or argue for the development of new ones. These conclusions have implications for further academic research. Given the theoretical nature of this thesis, theoretical conclusions and implications for further research are more prevalent, although most chapters also have policy implications.

The most important conclusion, which is important both for further academic research and for policy, is that the industrial organization of transport markets matters. The technical chapters above have each analyzed a specific competitive setting, but in each case, this competitive setting turned out to be important. Transport operators can use price discrimination to increase their profits, and this has implications for social welfare. In a multi-modal network, the fact that users can change their departure times affects the nature of competition, and the desirability of various competitive structures. Operators also use their departure times strategically, and change their investment strategies depending on the competitiveness of the market. Hence, when analyzing transport markets, it is crucially important to decide if, and how, the industrial organization of the markets will be modeled. In many cases, assuming that markets are perfectly competitive, or monopolistic with a perfectly elastic demand, such that market participants are still price takers, may not be appropriate. Similarly, when trying to address a policy issue, it is important to consider the full complexity of the specific transport market in question. Previous analyses of other markets, with different levels of competition, different demand structures, or other, perhaps on first sight small differences, will not always be helpful.

More specific conclusions can be drawn from the individual chapters of this thesis. Chapter 2 analyzed price discrimination and price differentiation in transport networks. This analysis has shown that it is important to make the distinction between differentiation and discrimination in a transportation context, especially if users have different values of time, or marginal costs differ for other reasons. Enforcing uniform or non-discriminatory pricing policies may improve social welfare in some cases, but decrease welfare in others. Hence, the stylized models do highlight the need for careful, situation-based analysis to evaluate the potential benefits of restrictive pricing policies. Importantly, policies that may be considered to be the most ‘fair’ by users, such as a ban on discriminatory pricing, are not always better for society as a whole than others such as a ban on price differentiation, even if users have different marginal costs.

Apart from these policy conclusions, chapter 2 has implications for further modeling, too. As it has shown, network structures can have a big impact on the model outcomes. In a network with serial links, a monopolistic operator may be able to discriminate based on users’ origins and destinations. This is not qualitatively different from a single-link network
where discrimination is only based on values of time or marginal external costs. There are quantitative differences, though, which means that a model with link-based pricing may not approximate a real-world situation well if, in reality, pricing is OD-based. In a network with parallel links, operators may discriminate based on route choices, and this is both qualitatively and quantitatively different than OD-based pricing.

Chapter 3 looked into the effect of travel time dynamics on competitive market outcomes, in a setting where an unpriced road ran parallel to a priced railway line. As shown in this chapter, the dynamic nature of the problem causes serial competition to influence fares, even in the absence of elastic demand. This results from the fact that, in a dynamic model, commuters have possibilities for intertemporal substitution, even if they do not have the option to not travel. Therefore, a fare change on one rail link changes not only the number of rail commuters relative to the number of road commuters, but also their trip timing decisions. If demand is fixed, the number of rail passengers will not necessarily affect anyone traveling on other links in the network, as was shown with a simple static model, but the changes in trip timing will change passengers flows in the whole network and through those, other operators’ patronages and profits. Naturally, a monopolistic operator internalizes the effect a price increase on one link has on the patronage of the other links, but a Bertrand-Nash operator disregards this, which leads to different, and potentially higher fares. For transport policy, this implies that, even in markets where demand is relatively price-insensitive, it may still be bad to have separate operators owning transport links that are complementary for most users.

Again, this chapter also has theoretical and practical implications for further research. Under the assumption that fares are time-invariant, it has proposed an alternative formulation of the problem, such that rail operators optimize their fares and the boundaries of arrival time intervals, rather than the commuter flows. Using these variables, closed form solutions can be obtained, which makes subsequent calculations considerably easier, although general analytical solutions are still difficult to obtain. The chapter has also shown that travel time dynamics are important in analyzing the benefits or costs of competition; hence, the decision to either use a dynamic model or a much simpler static model needs to be taken carefully.

Chapter 4 proposed a methodology to model the scheduling decisions of competing transport operators, using a Hotelling horizontal differentiation model, generalized to include price-sensitive demand and asymmetric schedule delay costs. This asymmetry represented the fact that it is usually more costly for travelers to be late than to be early. As was shown in the chapter, transport operators can schedule their services in a strategic manner, to avoid head-on competition. This implies that, next to fares and frequencies, departure times should also interest public transport regulators. If the socially optimal departure times and fares are not attainable, regulating one of these two variables can result in a modest efficiency improvement; travelers’ values of schedule delay determine which of the two results in the greatest gain. The chapter also concluded that regulation can create equilibria in situations where an unregulated market fails to do so. This suggests an additional benefit of regulation in this particular market: it can improve welfare, relative to competitive equilibria, but it can also create stability in situations where competitive equilibria do not emerge naturally.

Chapter 4 also has implications for modeling, even beyond a transportation context. The models that were developed generally have stable equilibria, which are interior: they do not necessarily result in minimum differentiation, and never in maximum differentiation. This is an attractive property, as these types of interior equilibria can be observed in many real-world markets. Moreover, the chapter has also shown that asymmetric distance costs
7.1 Results and implications

can be included in horizontal differentiation models. These asymmetries are also present in other settings, such as in the scheduling of TV programs, or in contexts where producers can differentiate their products in a quality dimension. Hence, this result opens up possibilities for applications in other markets. Asymmetric transport costs do result in asymmetric equilibria, with producers charging different prices, and hence, determining where these equilibria lie is much more involved. For some parameters, no stable equilibria may exist, due to the presence of possibilities for profitable undercutting.

Chapter 5 compared several methods to determine the stochastic user equilibrium (SUE) traffic flows through a congested network. This chapter is more methodological than the preceding chapters, and it therefore has fewer direct policy implications. Indirectly, however, its implications are important. As shown in this chapter, simple logit stochastic user equilibrium models can give results that are very similar to more general probit models. This result stems mostly from the fact that transport networks are congestible, which implies that the systematic utilities that users derive from overlapping routes are correlated, even if the random utilities corresponding to the routes are not. Moreover, in networks, OD-level differences can potentially cancel each other out. The differences in link flows between models are not systematic, i.e., one model does not always result in higher profit-maximizing or socially optimal tolls than another; this depends on the characteristics of the network. Since logit models are much less computationally intensive, this indicates that they might, at least in congested settings, be a better choice. As logit models need no simulation for the calculation of choice probabilities, they can lead to more accurate results and allow for studying of more and more complex policy instruments and games (e.g., tax competition, or networks with multiple operators). However, they do have one important caveat. SUE models can easily be overcalibrated, for instance, through the introduction of alternative-specific constants that have little theoretical justification. These overcalibrated models will fit the existing data better, but, when they are used to evaluate the effects of changes in model parameters, they can lead to large differences between the various SUE specifications. Careful calibration is therefore always necessary, and the fit of the model with existing data should not be the main, or at least not the only criterion.

Finally, chapter 6 proposed a two-stage methodology to capture the strategic formation of transport networks by one or more private operators. The models developed in this chapter can be used to analyze settings where transport networks are not designed by a single actor, but emerge as a result of competition. The chapter examined several settings: an unrestricted monopolistic or competitive market, a market where only fares or networks are regulated, and a fully regulated setting. In each case, a stochastic user equilibrium model determines how many users travel, and which routes they will take through the network. These models are difficult to solve, but, in some cases, integer programming or a smart semi-exhaustive search can go a long way. From a policy perspective, the most important conclusion that can be drawn from this chapter is that, even if both the the total amount of capacity and the fares can be regulated, first-best outcomes still cannot always be achieved. Even if the total amount of capacity is optimal, this capacity will not necessarily be located in the correct part of the network: allocative inefficiencies remain. These allocative inefficiencies are likely to be small, since private operators do not have incentives to construct a network that is very different from the optimum. In simple numerical models, the same holds in a competitive market, where multiple operators own links; this, however, does not generalize to larger networks, where multiple equilibrium network configurations may exist, some of which are likely to be much further away from the optimum configuration. If only fares or only network structure
can be regulated, the model parameters determine which of the two strategies is best.

7.2 Avenues for future research

This book has explored the industrial organization of transport markets. Naturally, it is impossible to answer all possible questions about such a broad subject in just one book. Moreover, for every question that is answered, several new ones usually arise; this is definitely true for what has been discussed above. Of the many avenues for future research, the following three are especially worth mentioning.

Theoretical work

First, although the theoretical models developed in this book provide valuable insight into how competition in transport networks can be modeled, and what the results of these models mean for real-world problems, there is scope for the development of other theoretical models. In particular, multi-level models with more levels than those developed above could be very useful. Assuming that a regulator can directly influence the behavior of transport operators, who in turn directly interact with users, is, in most settings, unrealistic. In reality, there are usually various levels of local government, station operators, separate infrastructure owners, etc. Interactions between all those market participants may well have a big impact on the way the market should be regulated.

Another dimension that has not been explored in this book is the effect of uncertainty and risk aversion. In most transport markets, there is a large amount of uncertainty about short-term variables, such as travel times, and even more uncertainty about long-term variables such as long-term demand trends, fuel prices, investment costs, and the regulatory environment. This will influence the decision making process of all market participants, particularly if, as is usually the case, they are risk averse. A complication here might be that the various market participants do not all have the same level of risk aversion: consumers may, for instance, be less risk averse than infrastructure investors or regulators. This will make the modeling of the market much more complex, but the results may also be very informative.

This book has assumed that regulators are synonymous with governments, and simply maximize social welfare. In reality, regulators are separate entities, and their objective is not necessarily the maximization of social welfare. They may, instead, minimize the cost of regulation, or the probability of a particular outcome. This has, so far, not received a lot of attention in the transportation literature. The behavior of regulators is, however, crucial, so there is a clear need for further analysis there.

Applied work

Secondly, there is a need for more applied work. Theoretical models, such as those developed in this book, are certainly useful. However, when, for instance, an actual regulator needs to determine what its optimal strategy is in a particular situation, these models will usually need to be applied to the real-world market in question. In the case of transportation markets, this is not trivial. Large-scale traffic assignment models are straightforward to build, and have indeed been developed for many real-world transportation networks. Including more realistic market representations, such as transport operators that compete with each other through prices and investment levels, is much more difficult.
Many of the models developed in this book are very computationally intensive, despite the small size of the networks, and the many simplifying assumptions. More powerful computers would be able to solve larger, more complex problems, but there is also a need for the development of new models that can be solved efficiently for large networks. Unfortunately, multi-level nonlinear models are notoriously difficult to solve. Advances in the operations research literature (importantly, the PATH solver (Dirkse et al., 2014), which can solve mixed complementarity problems, a common way to model imperfectly competitive markets with multiple levels) are creating new opportunities. Using the theoretical models developed in this thesis, it may be possible to analyze much more complex real-world markets. The results of these analyses would be of direct relevance for regulators, but can also inform the academic debate about the effects of competition in transport networks.

**Empirical work**

Finally, the models that have been developed in this book are mostly theoretical. It will be very interesting to see whether the theoretical conclusions that have been drawn from them can be backed up with more real-world evidence. Chapter 3 of this book has, for instance, argued that if travelers choose departure times, serial competition will affect profits and usage levels through departure time choices, not just through demand levels. There are transport networks where several operators own links that are used as complements (the Sydney Orbital Network is just one example), and hence, it may be possible to empirically verify whether the theory holds. Although this is outside the scope of this book, these types of empirical verification are important, especially in a field where there are often many competing assumptions or theories. Given the increasing levels of privatization that can currently be observed in many transport networks, the availability of data about the performance of various market structures should increase, which will create more opportunities for this type of work than there have been in the past.
Bibliography


Bibliography


Bibliography


Schrank, D., Eisele, B., Lomax, T., 2012. TTI's 2012 urban mobility report. Texas A&M Transportation Institute, College Station, TX.


Bibliography


Samenvatting (Dutch summary)

De industriële organisatie van transportmarkten: prijzen, investering en regulering in spoor- en weggennetwerken

Economen houden zich bezig met de distributie van schaarse middelen, zoals tijd of geld. Deze distributie vindt vaak plaats in markten: concrete of abstracte plaatsen waar consumenten en producenten elkaar ontmoeten. Markten zijn vooral interessant als ze niet de best mogelijke resultaten opleveren. ‘Best mogelijk’ kan, in dit geval, meerdere dingen betekenen, zoals een Pareto optimum (een situatie waarin niemand er op vooruit kan gaan zonder dat iemand anders er op achter uit gaat) of sociaal optimum (zodat de maatschappij als geheel er niet op vooruit kan gaan). In beide gevallen geeft een marktuitkomst die lager ligt dan het optimum overheden een reden voor interventie; de vraag is dan, uiteraard, welke interventie het meest geschikt is om het marktevenwicht terug te brengen naar het optimum.

Er bestaan meerdere situaties waarin markten niet optimaal functioneren; deze worden collectief ‘marktfalen’ genoemd. Een situatie die in transportmarkten vaak voorkomt is de aanwezigheid van externe effecten: kosten of baten van, in dit geval, reizen, die niet door de reiziger maar door anderen worden gedragen. De extra filedruk die iedere automobilist op andere automobilisten uitoefent is bijvoorbeeld een negatief extern effect. Externe effecten worden vaak in isolatie onderzocht, waarbij de aannome is dat er geen andere vormen van marktfalen bestaan. Veel studies nemen bijvoorbeeld aan dat markten perfect concurrerend zijn, zodat geen enkele producent of consument invloed kan uitoefenen op marktprijzen.

In de praktijk kunnen zowel consumenten als producenten marktprijzen vaak wel in hun voordeel beïnvloeden: monopolisten kunnen meer winst behalen door hun prijzen te verhogen als de vraag naar hun product niet perfect elastisch is, en oligopolistische producenten beheersen een dusdanig groot gedeelte van de markt dat ook zij invloed kunnen uitoefenen op prijzen. Dit kan leiden tot marktfalen, en in veel situaties kunnen deze effecten niet worden genegeerd. Deze dissertatie bestudeert daarom de industriële organisatie van transportmarkten: dat wil zeggen, de beslissingen die worden genomen door economische agenten in markten die niet volkomen concurrerend zijn, de implicaties van die beslissingen, en de manieren waarop overheden ze kunnen beïnvloeden. In deze dissertatie worden methodes gepresenteerd die gebruikt kunnen worden om zulke beslissingen te modelleren. Deze methodes worden vervolgens toegepast om prijssstrategiën, de keuze van bus- en treintijden, investeringsbeslissingen en reguleringsstrategiën te analyseren. In contrast tot een groot deel van de bestaande literatuur wordt hierbij de netwerkstructuur van transportmarkten expliciet meegenomen.

Economische modellen zijn vereenvoudigde voorstellingen van de werkelijkheid, en maken altijd aannames. Het is zelden het geval dat een enkel model volstaat om een complexe markt te analyseren; het is vaak nodig verschillende modellen te gebruiken, die verschillende aannames maken en gebruikt kunnen worden om verschillende aspecten van de markt te belichten. Deze dissertatie bestaat daarom uit vijf aparte hoofdstukken, naast de introductie en conclusie. Ieder hoofdstuk belicht een belangrijk aspect van een of meerdere transport-
Samenvatting (Dutch summary)

Ieder hoofdstuk heeft een dubbel doel: eerst worden nieuwe methodes ontwikkeld om de industriële organisatie van transportmarkten te bestuderen, en vervolgens worden deze methodes in de praktijk gebracht om conclusies te kunnen trekken over mogelijke beleidsinterventies.

Hoofdstuk 2 gaat in op de effecten van prijsdifferentiatie en prijsdiscriminatie door een monopolistische vervoerder, die opereert in een netwerk waarin congestie kan optreden. In drie verschillende ruimtelijke modellen wordt nagegaan wat het effect is van het verbieden van prijsdifferentiatie (zodat alle reizigers voor het zelfde reisproduct het zelfde betalen) en/of prijsdiscriminatie (zodat prijzen alleen mogen variëren met de marginale externe kosten van reizigers). Doordat de drie modellen verschillende ruimtelijke structuren hebben kunnen prijsdiscriminatie en prijsdifferentiatie op basis van gebruikersgroep, route, en herkomst en bestemming apart worden geanalyseerd. Dit hoofdstuk generaliseert de bestaande literatuur over prijsdiscriminatie, waarin gebruikersgroepen doorgaans alleen verschillende tijdswaarderingen hebben, door toe te laten dat groepen ook verschillende marginale externe kosten hebben. Prijsdiscriminatie en prijsdifferentiatie zijn daardoor niet langer hetzelfde. Het hoofdstuk laat zien dat het verbieden van prijsdiscriminatie en prijsdifferentiatie soms welvaartswinst kan opleveren, maar soms ook tot welvaartsverlies leidt. Het verbieden van prijsdifferentiatie kan beter zijn dan het verbieden van prijsdiscriminatie, ondanks het feit dat die laatste optie als eerlijker gezien kan worden.

Hoofdstuk 3 onderzoekt netwerken waarin meerdere oligopolistische vervoerders actief zijn. Het analyseert de strategische beslissingen van marktteilnemers in netwerken waarin naast openbaar vervoer ook onbeprijsde wegen bestaan, en waarin sommige gebruikers meerdere treinen of bussen nodig hebben om hun bestemming te bereiken. Het hoofdstuk contrasteert twee modellen: een waarin een monopolistische aanbieder al het openbaar vervoer bezit, en een waarin iedere individuele verbinding door een aparte aanbieder wordt geëxploiteerd. Gebruikers kiezen niet alleen hoe ze reizen, maar ook wanneer; de modellen zijn dynamisch. Deze dynamiek is belangrijk, zoals het hoofdstuk laat zien. In tegenstelling tot wat in de bestaande, op statische modellen gebaseerde literatuur te zien is, kan in de modellen die in dit hoofdstuk gepresenteerd worden seriële concurrentie prijzen verlagen, in plaats van verhogen.

Hoofdstuk 4 heeft, net als hoofdstuk 3, betrekking op reistijdstipkeuzes. In dit hoofdstuk wordt specifiek ingegaan op de manier waarop oligopolistische vervoerders hun vertrektijden bepalen, en vooral op hoe deze vertrektijden strategisch gebruikt kunnen worden. Om deze vragen te kunnen beantwoorden ontwikkelt dit hoofdstuk een gegeeneraliseerde versie van Hotelling’s model van horizontale differentiatie, waarin de vraag naar vervoer prijsafhankelijk is, en waar afstands kosten asymmetrisch zijn. In dit model kiezen twee concurrenten prijzen en vertrektijden binnen een vast tijdsinterval; reizigers zijn uniform verdeeld over dit interval, en de locatie van een reiziger geeft de vertrektijd aan die zijn of haar voorkeur heeft. Dit hoofdstuk laat zien dat vertrektijden strategisch gebruikt kunnen worden om winst te maximaliseren, en bespreekt welke strategieën het beste gebruikt kunnen worden om de markt te reguleren.

Hoofdstuk 5 wordt de routekeuze van reizigers behandeld. In de routekeuzeliteratuur worden verschillende modellen voorgesteld; dit hoofdstuk onderzoekt hoe groot de verschillen tussen drie van deze modellen (multinomiaal probit, nested logit, and generalized nested logit) zijn. Om dit te kunnen bepalen worden de modellen toegepast op een klein maar representatief netwerk waarin congestie kan ontstaan. Ze worden vervolgens gebruikt om beleidskeuzes te evalueren. Zoals dit hoofdstuk laat zien zijn de implicaties van de keuze voor een bepaald model voor beleidskeuzes klein, mits de modellen goed gecalibreerd zijn. Het is
daarom verdedigbaar om simpelere modellen, zoals het nested logit of generalized nested logit model te gebruiken, in plaats van meer realistische maar ook meer computationeel intensieve modellen zoals het probit model.

De resultaten van hoofdstuk 5 worden gebruikt in hoofdstuk 6. In dit hoofdstuk wordt een generalized nested logit model gebruikt als de basis voor een lange-termijn investeringsmodel, waarin private oligopolistische vervoerders zels nieuwe verbindingen kunnen aanleggen. Dit model wordt vervolgens gebruikt voor de evaluatie van beleidsinterventies die als doel hebben de formatie van transportnetwerken te reguleren. In tegenstelling tot de meeste bestaande literatuur zijn investeringen in nieuwe vervoerscapaciteit discreet: een verbinding wordt aangelegd of niet. Dit heeft belangrijke implicaties; vooral omdat, in dit geval, prijsbeleid niet langer afdoende is om het sociaal optimum te herstellen.

Samen geven deze hoofdstukken een breed overzicht van de verschillende kwesties die betrekking hebben op de industriële organisatie van transportmarkten. De belangrijkste conclusie die uit deze dissertatie getrokken kan worden is dat deze industriële organisatie belangrijk is. De verschillende hoofdstukken behandelen elk een specifieke marktvorm, en in ieder hoofdstuk blijkt dat het modelleren van het strategische gedrag van vervoerders en reizigers cruciaal is. Vervoerders kunnen prijsdiscriminatie en prijsdifferentiatie toepassen om meer winst te behalen, en dat heeft implicaties voor het maatschappelijk welzijn. In multimodale netwerken heeft de reistijdstipkenze van reizigers een belangrijk effect op de manier waarop vervoerders concurreren, en op de relative efficiëntie van verschillende marktvormen. Vervoerders kunnen ook hun vertrekttijden op een strategische manier gebruiken om zo meer winst te behalen, en ook dat heeft implicaties voor het maatschappelijk welzijn en daarmee ook voor beleid. Het is daarom bij elke analyse van transportmarkten van cruciaal belang om de industriële organisatie correct te modelleren, met de juiste hoeveelheid detail. In veel gevallen zal het niet afdoende zijn om simpelweg aan te nemen dat de markt volkomen concurrerend is. Hetzelfde geldt voor beleidsanalyses: ook daar is het belangrijk om de complexiteit van de markt niet te negeren. Bestaande beleidsanalyses van andere markten, die meer of minder competitief zijn, een andere vraagstructuur hebben, of die op een andere, wellicht op het eerste gezicht kleine manier verschillen zullen vaak niet afdoende zijn.
Samenvatting (Dutch summary)

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

562 T. SALIMANS, Essays in Likelihood-Based Computational Econometrics
563 P. SUN, Tail Risk of Equidity Returns
564 C.G.J. KARSTEN, The Law and Finance of M&A Contracts
565 C. OZGEN, Impacts of Immigration and Cultural Diversity on Innovation and Economic Growth
566 R.S. SCHOLTE, The interplay between early-life conditions, major events and health later in life
567 B.N. KRAMER, Why don’t they take a card? Essays on the demand for micro health insurance
568 M. KILIÇ, Fundamental Insights in Power Futures Prices
569 A.G.B. DE VRIES, Venture Capital: Relations with the Economy and Intellectual Property
570 E.M.F. VAN DEN BROEK, Keeping up Appearances
571 K.T. MOORE, A Tale of Risk: Essays on Financial Extremes
572 F.T. ZOUTMAN, A Symphony of Redistributive Instruments
573 M.J. GERRITSE, Policy Competition and the Spatial Economy
574 A. OPSCHOOR, Understanding Financial Market Volatility
575 R.R. VAN LOON, Tourism and the Economic Valuation of Cultural Heritage
576 I.L. LYUBIMOV, Essays on Political Economy and Economic Development
577 A.A.F. GERRITSEN, Essays in Optimal Government Policy
578 M.L. SCHOLTUS, The Impact of High-Frequency Trading on Financial Markets
579 E. RAVIV, Forecasting Financial and Macroeconomic Variables: Shrinkage, Dimension reduction, and Aggregation
580 J. TICHEM, Altruism, Conformism, and Incentives in the Workplace
581 E.S. HENDRIKS, Essays in Law and Economics
582 X. SHEN, Essays on Empirical Asset Pricing
583 L.T. GATAREK, Econometric Contributions to Financial Trading, Hedging and Risk Measurement
584 X. LI, Temporary Price Deviation, Limited Attention and Information Acquisition in the Stock Market
585 Y. DAI, Efficiency in Corporate Takeovers
586 S.L. VAN DER STER, Approximate feasibility in real-time scheduling: Speeding up in order to meet deadlines
587 A. SELIM, An Examination of Uncertainty from a Psychological and Economic Viewpoint
588 B.Z. YUESHEN, Frictions in Modern Financial Markets and the Implications for Market Quality
589 D. VAN DOLDER, Game Shows, Gambles, and Economic Behavior
590 S.P. CEYHAN, Essays on Bayesian Analysis of Time Varying Economic Patterns
591 S. RENES, Never the Single Measure
D.L. IN 'T VELD, Complex Systems in Financial Economics: Applications to Interbank and Stock Markets

Y.YANG, Laboratory Tests of Theories of Strategic Interaction

M.P. WOJTOWICZ, Pricing Credits Derivatives and Credit Securitization

R.S. SAYAG, Communication and Learning in Decision Making

S.L. BLAUW, Well-to-do or doing well? Empirical studies of wellbeing and development

T.A. MAKAREWICZ, Learning to Forecast: Genetic Algorithms and Experiments

P. ROBALO, Understanding Political Behavior: Essays in Experimental Political Economy

R. ZOUTENBIER, Work Motivation and Incentives in the Public Sector

M.B.W. KOBUS, Economic Studies on Public Facility use

R.J.D. POTTER VAN LOON, Modeling non-standard financial decision making

G. MESTERS, Essays on Nonlinear Panel Time Series Models

D. KOPÁNYI, Bounded Rationality and Learning in Market Competition

N. MARTYNOVA, Incentives and Regulation in Banking

D. KARSTANJE, Unraveling Dimensions: Commodity Futures Curves and Equity Liquidity

T.C.A.P. GOSENS, The Value of Recreational Areas in Urban Regions

L.M. MARČ, The Impact of Aid on Total Government Expenditures

C.LI, Hitchhiking on the Road of Decision Making under Uncertainty

L. ROSENDAHL HUBER, Entrepreneurship, Teams and Sustainability: a Series of Field Experiments

X. YANG, Essays on High Frequency Financial Econometrics