Title
Spatial Models of Morning Commute Consistent with Realistic Traffic Behavior

Author
Lago, Alejandro

Publication Date
2003

Institute of Transportation Studies
University of California at Berkeley

Spatial Models of Morning Commute Consistent with Realistic Traffic Behavior

Alejandro Lago

DISSERTATION SERIES
UCB-ITS-DS-2003-1

Fall 2003
ISSN 0192 4109

Spatial Models of Morning Commute Consistent with Realistic Traffic Behavior

by

Alejandro Lago

Engineering (Technical University of Catalonia, Spain) 1998
M.S. (University of California at Berkeley) 1999

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering — Civil and Environmental Engineering

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Carlos F. Daganzo, Chair
Samer M. Madanat
John M. Quigley

Fall 2003

The dissertation of Alejandro Lago is approved:

Chair Date

Date

Date

University of California, Berkeley

Fall 2003

Abstract

Spatial Models of Morning Commute Consistent

with

Realistic Traffic Behavior

by

Alejandro Lago

Doctor of Philosophy in Engineering — Civil and Environmental Engineering

University of California, Berkeley

Carlos F. Daganzo, Chair

Urban planners are increasingly concerned about the sprawling suburban development in metropolitan areas around the world, which they often blame for growing traffic congestion and excessive highway investment needs. This dissertation seeks to shed light on this issue by studying the relationship between morning commute congestion and urban form.

The causes and consequences of traffic congestion have been extensively studied in the economics and engineering literatures. Unfortunately, most conclusions have been drawn from very idealized models, which either fail to consider adequately the spatial nature of congestion, by neglecting the effects of physical queues and merging interactions, or overlook dynamic aspects, such as commuters' departure time adaptation during the rush-hour.

To better capture the spatial-dynamic nature of morning commute traffic, this dissertation proposes a new analytical framework that explicitly incorporates *spatially distributed commuter origins*, *realistic traffic behavior* and *commuter timing*

decisions. The work combines the departure-time equilibrium theory (as first proposed by Vickrey [1969]) with the spatial model of traffic dynamics of Newell [1993] and the model of merge traffic interactions of Daganzo [1994, 1995a].

Focus is placed on idealized urban configurations, where traffic behavior can be studied analytically and general insights can be gained. We first study the equilibrium problem in a stylized two-origin network. This enables us to understand the fundamental role of merging bottlenecks and queue spillovers when commuters have different origins. The analysis is then extended to model congestion behavior in long freeway corridors and monocentric cities. We develop an exact procedure to solve the dynamic departure-time equilibrium for single-destination freeway tree networks. Solutions are characterized for cases with and without an alternative street network. The results show that the location-based congestion cost is very dependent on the spatial behavior of queues and that congestion can be reduced by altering the freeway access priorities given to different origins. At the same time, urban sprawl is shown to contribute not only to larger travelled distances but also to increased overall delays. Sprawl effects, however, are not as severe as often assumed.

We finally propose some closed-form continuous approximations for the location-based congestion cost. These formulae provide an improved and simple representation of the dependence of congestion on the spatial distribution of population that can be easily incorporated to study policy issues. The design of more effective measures to reduce congestion and control urban development is an immediate example.

Carlos F. Daganzo
Dissertation Committee Chair

To my parents, Luis and Manuela

This dissertation is mainly the result of their stubbornness

. . . 'Cos I know what it means
To walk along the lonely street of dreams . . . .

— WHITESNAKE, *Here I go again* (1987)

Contents

List of Figures

List of Tables

Acknowledgments

For many of you accidental readers, the acknowledgements may be the first and only part you will ever read of this piece of work of mine. I hope at least to convey to you what a worthy effort writing this thesis was, even if only for the sake of having the chance to thank all the people that played a part in it.

I feel most indebted to my friend, mentor and advisor – strictly enforced by this order – Carlos Daganzo. He not only provided me with knowledge and support but also the most fundamental value I always needed: confidence. He is undoubtedly the only person on Earth that made me feel ashamed . . . by often taking more interest in this dissertation than I did.

My gratitude is extended to the extraordinary pool of professors with whom I had the chance to engage during my time in Berkeley, especially, Samer Madanat who taught me in the art of teaching and shared with me a wide variety of interests, from politics to soccer; Adib Kanafani, whose support during my early years in Berkeley provided me with a broader perspective to face many problems; John Quigley, who willingly offered his time and knowledge on urban economics to complement my work in this thesis; and also, Francesc Robusté, transportation professor in Barcelona, whose encouragement and advice made my inner 'Hamlet' finally decide to come to Berkeley.

Just as invaluable to me has been the help of Mari Mordecai, Cindy Kennon and the rest of staff at the Civil Engineering Academics Office and ITS. I can only hope that for the sake of their mental well-being other students are better organized than me. The same can be said about the ITS Harmer Davis Library staff. To all of them, I clearly owe part of the P, h and D just as this magnificent institution - the University of California - owes part of its renowned excellence.

Obviously, academics is only a minuscule part of what the Berkeley Circus fed

into my life and I truly feel Berkeley has granted me a Ph.D. in friendship and life experience. Since I could only hope to refer to an appendix to include all the people I had the opportunity to share my life with, I will at least mention here my roommates Asim, Julian 'Pelaíto' and Sean (and of course, our virtual roommate Muge), my spanish fellows Chema, Ruben, Guillermo, Alfonso, Erika and Rafa in representation of the Iberia-Berkeley community, and my transportation pals Tanja Bolic, Irwin Guada, Yuwei Li, Sebastian Renaud, Aaron Golub and Matt Malchow, of which the latter two hold the more-than-honorable title of 'American' friends. Above all, nevertheless, I need mentioning Juan Carlos Muñoz, who always held the flag I needed to follow during all these years.

Finally, I can only express my deepest love for my family: my parents, to whom this thesis is dedicated since, I insist, this work is more the result of their stubbornness than of my own effort; my grandparents, to whom my absence will never pay off; my brother Hector, somebody who may sometimes feel forgotten but who has taught me that role models also work in the younger-to-older brother direction; and finally, Maria, once my love, now my best friend and always my 'pookie', the memory of Berkeley will remain indelibly tied to her in my heart.

Final note: I am also grateful to the University of California Transportation Center (UCTC) for the financial support throughout my last year of research.

Chapter 1

Introduction

THIS DISSERTATION seeks to shed light on the relationship between morning commute congestion, commuter trip timing and urban form. Understanding how congestion is generated is a necessary first step towards the design of effective and equitable measures to reduce congestion.

Traffic congestion ranks among the top problems in metropolitan areas [UNDP, 1997].[1] The traditional approach to mitigate congestion has been the construction or expansion of road infrastructure, but this approach seems no longer viable given the growing scarcity of land and public funds. As a result, the interest of urban planners and politicians has gradually gravitated towards devising solutions that will manage the existing demand more efficiently. These solutions range from short-term policies aimed at controlling day-to-day traffic flows (i.e., mainly through road pricing or some other strategy to control access to the road network) to more long-term strategies aimed at redirecting land-use patterns. Transportation practitioners, however, seldom

[1] Although congestion estimates should be regarded with caution, statistics on urban travel suggest that morning commute has worsened substantially on the last decades. For instance, traffic delays per person resulting from traffic congestion increased by more than 200 percent from 1982 to 2000; see TTI [2002].

agree on the plausible outcomes of many of these policies. Not surprisingly, few go beyond the drawing board since politicians prefer to avoid the substantial risk that their implementation (even if done on an experimental basis) entails. This lack of consensus is best exemplified by the current debate on urban sprawl and the long-term policies necessary to achieve a sustainable urbanization of our cities. Urban areas continuously expand as people prefer the affordable suburbs to the dense, and often more expensive, central urban locations. Some point to this unstoppable urban sprawl as the main cause of congestion and call for measures to control it (e.g., mixed-use land development, land-use taxation and other accessibility enhancing measures); opponents, on the other hand, content that this *smart growth* approach can only lead to denser cities with more congestion and pollution (see Cox [2000]). A better understanding of the mechanisms that link congestion and urban form is hence needed to guide this policy debate.

Unfortunately, although researchers have devoted substantial effort in finding adequate ways to describe and analyze congestion for more than 75 years, not many clear and reliable insights have arisen. The analysis of the mechanisms that generate congestion is indeed a complex task. Traffic congestion is not just the result of the total volume of trips done on a given metropolitan area, as the traditional economic representation often assumes, but of the way the trips take place in space and time. Commute trips are spatially distributed and, as such, congestion levels change with location. Some locations concentrate a major number of trips and arise as natural bottlenecks in the network. Queues at these locations spread over the network thereby affecting locations (and commuters) differently. Traffic is also a dynamic phenomenon. Congestion levels change substantially during rush-hour and people, particularly during the morning commute, tend to adapt the routing and scheduling

of their departure times to respond to the varying congestion. A robust characterization of traffic behavior, even at the macroscopic level needed for policy analysis, must recognize all these interactions explicitly.

This mixed spatial-dynamic nature of the traffic phenomena poses important challenges in terms of both the mathematical representation of the problem and its analysis. To cope with these difficulties, research has been pulled to two different extremes, as evidenced by the two main literatures on the field. Engineering-style analysis has largely drawn from very complicated simulation/optimization models that require assumptions not always fully realistic and whose results are often too particular to the model assumed or too cumbersome to draw general insights for policy analysis. On the other hand, economic/planning-style analysis has been largely based on very stylized models that overlook many of the relevant margins of analysis (e.g., dynamic issues like trip timing are neglected and/or traffic models are often assumed that treat incorrectly the interaction of trips in space).

The research on this dissertation sits on the middle-point of these two approaches. We propose a new framework to analyze the temporal and spatial interactions by considering models of morning commute that jointly incorporate *spatially distributed commuter origins*, *realistic traffic behavior* and *commuter timing decisions*. Crucial questions that are answered include: (1) how does congestion develop in cities as a function of the spatial distribution of population? (2) how do congestion costs suffered by commuters differ by location? and (3) how do these costs may affect commute travel decisions? Because we seek to obtain concrete qualitative results that will guide policy more effectively, focus is placed on models that can be solved analytically and from which general insights can be drawn. Idealized urban configurations are used for that purpose. More importantly, we seek to provide a versatile (i.e., analytical)

tool to assess how costs and commuter decisions may be affected by different policies.

The remainder of this chapter frames the work in this dissertation within the existing research. In section 1.1 we review the state-of-the-art congestion analysis. Section 1.2 motivates the scope of this research, summarizes the main contributions and outlines the organization on the thesis.

1.1 Traffic congestion analysis: state-of-the-art

Research on traffic congestion encompasses work in four main related areas: traffic theory, network equilibrium modelling, road pricing and investment analysis, and urban location theory. The first two areas are directly concerned with the prediction of traffic behavior and are predominantly the domain of transportation engineers and scientists. The latter two areas concentrate on the economic effects of congestion, particularly on how to achieve an efficient use of the road infrastructure using control mechanisms and on how land use patterns and congestion are related.

Obviously a full review of a research area so broad in scope cannot be presented in an introductory chapter. Therefore, we will only focus on the historical evolution of congestion studies, highlighting the main approaches and discussing their limitations. Our objective is to identify the areas where improvement is necessary.

1.1.1 Traffic theory

Traffic models are fundamental in any analysis of congestion. They represent the physics governing the interaction of vehicles in a traffic stream and provide formulations that allow estimating travel conditions as a function of known vehicle volumes and road characteristics. Models can be coarsely divided in two main groups: micro-

scopic models (i.e., those that study individual vehicle interactions) and macroscopic models (i.e., those that represent traffic as fluid and study only the behavior of some aggregated variables). We limit our attention to the latter models since those are the relevant ones for policy and economic analysis.

A significant portion of the literature focuses on steady-state (or time invariant) representations of traffic. Stationary traffic conditions are represented by a *fundamental diagram*, which describes the possible states of a homogenous traffic stream in terms of three main parameters: density k (vehicles per unit length), speed v (distance per unit time) and flow q (vehicles per unit time). The fundamental diagram varies with the road characteristics, traffic composition and other environmental factors, but it has a basic shape represented in Figure 1.1a (see May [1990] for a literature review). As shown in the upper part of the figure, speed decreases monotonically with density (i.e., the number of vehicles in the road), with the decrease being steeper with larger densities when queues appear. This behavior is not surprising since commuters tend to keep smaller spacing between vehicles as speed decreases. Traffic flow is determined by speed and density through the identity $q = kv$; therefore, it first increases with density up to a maximum flow (called the *capacity* q_{max}) and then decreases until the density reaches a maximum value (called the *jam density* k_j) when both flow and speed are zero; see the lower part of Figure 1.1a. The increasing branch of the $q - k$ diagram is normally called the *free-flow* or *uncongested* regime, since the stationary states in this branch arise when no restrictions exist downstream of the road. The decreasing branch is called the *congested* or *queued* regime since its states describe the traffic stream inside queues caused by downstream restrictions. The decreasing branch is sometimes termed the *hypercongestion* regime in the economics literature and the term *volume* is also often used instead of flow.

For network design, planning and economic analysis, the usual goal is predicting travel times. Many studies customarily use, instead of the fundamental diagram, some form of a link-based *volume vs. trip time curve* for that purpose. Volume-trip time curves give the average vehicle trip time as an increasing function of the *volume/capacity ratio* under stationary conditions; see Figure 1.1b. Mathematically, these functions can be derived from the uncongested branch of the fundamental diagram since the trip time for a link of length ℓ is ℓ/v and the volume capacity ratio is $q/q_{\max}$. These curves are often called *link performance functions* or *volume-delay* curves. Because of their convenience, link performance functions have become a standard tool of the network literature (see section 1.1.2), and many *ad-hoc* forms (not necessary consistent with the traffic fundamental diagram) have been proposed and tested empirically; see Branston [1976].[2] There are, however, some important limitations that are not always clearly recognized in this representation of traffic. First, link performance functions represent steady-state behavior that cannot be assumed to hold during a full peak period. Hence, link performance functions estimated from time-dependent data may internalize not only the technological aspects of the road segments where they are estimated, but also features of its demand. Second, from a spatial point of view, these functions ignore spatial interactions between connected links (e.g., interactions at merges or diverges). Hence, their use on most network settings is unrealistic. Finally, volume-delays function can only represent adequately situations of mild congestion. When queues fill a link, experience and experiments show that longer travel times arise when flow declines, a result contrary to the link function prediction. Obviously, this limitation reduce the applicability of such functions since networks without spillovers are rare (i.e., the situations of interest are

[2]By far the most widely used link performance functions are the *BPR* curves [Bureau of Public Roads, 1964].

normally those where large delays are experienced by commuters). To overcome these limitations, traffic models must explicitly consider the traffic dynamics.

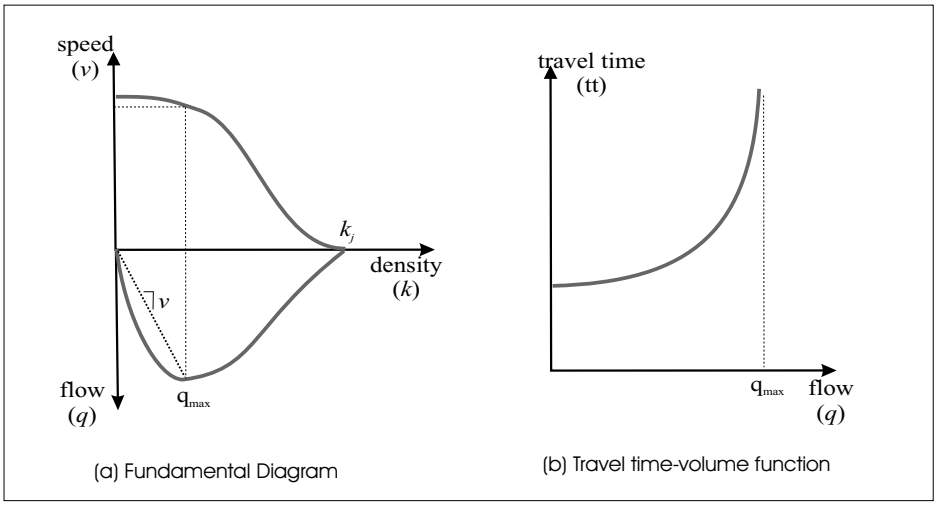

Figure 1.1. Steady-state traffic model.

Dynamic models allow for traffic conditions (e.g., flow, density and speed) to vary with location and time. The simplest and most widely accepted dynamic model is the *kinematic wave* (hereafter KW) model, first proposed by Lighthill and Whitman [1955] and Richards [1956]. This model assumes that traffic can be treated at the macroscopic level like a fluid and that the stationary fundamental *flow-density* diagram holds also under non-stationary conditions at every location and time. Transitions between different stationary states (e.g., between a free-flow and a queued situation) are represented by waves that propagate along the road segment. The KW theory – as originally presented by Lighthill and Whitman, and Richards – requires a burdensome mathematical apparatus to obtain solutions. This inconvenience has limited its use for practical applications, and several alternative approaches have been proposed. The most common simplification consists in assuming that the static volume-delay functions also apply to the dynamic case; i.e., that the travel time (or

for that matter, the speed) of a vehicle entering a road is only a function of the road inflow at the time of entry. This is equivalent to assuming that congestion is a local phenomena and that no propagation of traffic conditions occurs on time or space (from now on, we call this the *local congestion* assumption). This simplification leads to clearly inconsistent behavior – for example, the classical *Smeed's paradox* [Smeed, 1967] in which vehicles departing late in a low-flow cohort catch up a high-flow cohort that departed earlier, overtake them and arrive to the destination before them.[3] To alleviate this problem, link performance functions have been amended to allow the travel time of an entering vehicle to depend on the link outflows and/or the occupancies (see Ran and Boyce [1996], chapter 12). However, Daganzo [1995c] shows the inconsistencies persist in any model where travel time depends in any way on the inflows or outflows. A model that explicitly accounts for flow propagation and avoids Smeed's paradox is proposed in Mahmassani and Herman [1984]. In this model, density and speed change simultaneously and uniformly along the link with every change in the link inflow. Unfortunately, such assumption implies rather unrealistically that traffic conditions propagate in the forward direction instantly (i.e, that vehicle speed continues to be affected by following traffic) and this has undesirable effects too; see Newell [1988].

A more consistent treatment of traffic dynamics is obtained by representing each link as a bottleneck with a fixed capacity and a dimensionless (or *point*) queue forming upstream when link inflow exceeds capacity; see, for instance, Kuwahara and Newell [1987] and Kuwahara and Akamatsu [1997]. With this assumption, link travel time depends exclusively on link occupancy at the time of entry. Although these models do not suffer from Smeed's paradox, they still yield wrong predictions when

[3] Newell [1988] showed that the paradox never arises if the speed of vehicles is affected by congested conditions ahead of them, as in the KW model.

queues spillover across links; see Daganzo and Lin [1994]. Thus, even queuing models fail to represent adequately the macroscopic traffic behavior under heavily congested situations.

Fortunately, Newell [1993] recently showed that the KW solution procedures can be simplified dramatically if (i) traffic is represented in terms of cumulative vehicle counts (instead of flows) and, (ii) a triangular fundamental q-k diagram is used. Under these assumptions, an alternative treatment based on standard queuing theory methods is possible, opening the door for the analytical treatment of some important problems. Daganzo [1994, 1995a] extend Newell's ideas by including consistent models of merge and diverge interaction and an efficient approximation procedure for very large networks (the so-called *cell transmission model*). Recent empirical evidence [Windover, 1998] has confirmed that the KW model captures the macroscopic behavior of queues quite realistically and provides estimates of overall vehicle delays in agreement with empirical observations. The KW model is known to have some limitations, but higher-order modifications to the KW model or alternative microscopic models based on *car following theory* do not necessarily present a better grounded representation of traffic [Daganzo, 1997] and have failed to provide better predictions of travel times [Brockfeld et al., 2003].

Newell KW procedure will be adopted throughout this dissertation. A detailed explanation of the theory is given in chapters 2 and 3.

1.1.2 Network models and equilibrium

Network models are used to simulate traffic behavior when different origins and destinations are linked through a series of routes. These models must recognize that vehicle flows on each route are not known *a priori*, since they are the result

of commuters' trip decisions. Commuters respond to congestion by choosing among different modes, routes and departure times. Since the focus of this thesis in on road-traffic congestion, we shall restrict this review to models that incorporate route and departure time choice.

The conceptual framework to analyze route choice under steady-state conditions was introduced by Wardrop [1952]. According to *Wardrop first principle*, users choose routes so as to minimize their individual trip cost. This behavior leads to an equilibrium situation in which the trip costs between each origin destination (OD) pair are equal in all used routes and larger on the routes not used; i.e., commuters do not have an incentive to change routes. The resulting pattern is normally called the *user equilibrium*. The user equilibrium differs from *system optimum* patterns where the minimum total travel time in the system is achieved as if an overseeing authority could direct all commuters. Wardrop's principle is an oversimplification of reality because it assumes perfect information and utility-maximizing commuters that behave deterministically,[4] but it is reasonable as a first approximation for rough planning analysis. Beckmann et al. [1956] proposed a mathematical optimization framework to solve Wardrop's network equilibrium problem under static traffic conditions using link performance functions. Extensions of Wardrop's principle to stochastic route choice have been provided in Daganzo and Sheffi [1977]. Extensions that consider multiple vehicle classes can be found in Dafermos [1980] and Daganzo [1983].[5]

Since the static models are not very satisfactory to represent situations of high congestion as mentioned in §1.1.1, dynamic traffic equilibrium has been an active area of research for the last two decades. Dynamic network equilibrium is both conceptually and computationally more difficult. A first conceptual complication stems from

[4]Wardrop principle is a special case of Nash equilibrium.

[5]For a more through review of static network models see Patriksson [1994].

the fact that different definitions of equilibrium are possible under dynamic conditions, depending on the information available to the users. The natural generalization of Wardrop principle assumes that the routes chosen on each OD pair at each time of departure are those that minimize the experienced trip cost (which is now time-dependent). This type of equilibrium, normally termed *ideal* or *predictive* dynamic user equilibrium ($PDUE$), assumes that the dynamic evolution of traffic conditions is consistent day-after-day and therefore, that users can be aware (i.e., predict) of future traffic conditions at links visited downstream when selecting optimal routes at their origin. Smith [1993], Wie et al. [1995], Ran et al. [1996], Ran and Boyce [1996], Akamatsu and Kuwahara [1999], Tong and Wong [2000], Akamatsu [2001] and Huang and Lam [2002] provide different network models under this assumption. Since equilibrium route choice must be based on future traffic conditions this type of equilibrium problems are notoriously difficult to solve. Equilibrium is generally modelled through a set of discrete variational inequalities [Wie et al., 1995] or an equivalent non-linear optimization problem [Akamatsu, 2001; Ran et al., 1996]. In all cases, laborious numerical search procedures (e.g., *Frank-Wolf*-like decomposition) are required to solve the problem. Furthermore, route enumeration is normally unavoidable (when multiple destinations exist) since it is necessary to keep track of flow propagation along specific routes. Therefore, substantial computational effort is required even for medium size problems. Alternatively, simulation-based approaches can be used [Huang and Lam, 2002; Smith, 1993; Tong and Wong, 2000] but still some sort of heuristic is needed to update volumes in each path until equilibrium conditions are approximately met.

An alternative representation of equilibrium, normally termed *instantaneous* or *reactive* dynamic user equilibrium ($RDUE$), assumes that commuters choose routes

at each time based on the current travel times prevailing on the network. Works in this class include Friesz et al. [1989] (extended in Wie et al. [1990]), Papageorgiou [1990], Janson [1991], Ran et al. [1993], Lam and Huang [1995] and Kuwahara and Akamatsu [1997]. This type of equilibrium is easier to solve since no prediction of future travel conditions is needed. For example, if one ignores the multi-commodity contraints, as done in Wie et al. [1990] or Papageorgiou [1990], the problem can then be formulated as a standard optimal control problem. A better approach solves a shortest path problem in each time interval [Kuwahara and Akamatsu, 1997]. Furthermore, route enumeration is not required. The assumption about instantaneous travel times, however, is not very realistic when commuters trip times are comparable in duration with the length of the rush.

In a dynamic traffic setting, commuters also reschedule their departure times based on congestion levels. Anecdotal evidence suggests that this scheduling adaptation may have effects as important as route choice;[6] notwithstanding, dynamic equilibrium models that explicitly consider commuter departure time choice have received less attention, perhaps because departure time equilibrium is more difficult to model. A framework to analyze departure-time choice was first proposed by Vickrey [1969]. Vickrey assumed that commuters have a preferred time of arrival to their destinations and schedule their departure (or arrival) times to avoid periods of high congestion at the expense of suffering a scheduled delay for arriving earlier or later than desired to their destination. Small [1982] provides empirical verification of this behavior. Vickrey analyzed equilibrium in a very simplified time-dependent scenario with a single bottleneck, a single destination and a fixed number of commuters (more details

[6]The omission of timing changes can lead to incorrect predictions about the benefits of policy measures such as congestion pricing or capacity expansions. Small [1992] discusses the example of BART opening in San Francisco.

are given in chapter 2). Vickrey's framework has been incorporated into network models with route choice in Kuwahara and Newell [1987], Bernstein et al. [1993], Wie et al. [1995], Ran et al. [1996] and Huang and Lam [2002]. These models inherit all the difficulties of route-choice *PDUE* models and need to be solved with heuristic and/or simulation-based techniques.[7]

A common limitation to all the equilibrium models above, however, is that they are based on traffic models that are not fully consistent. The models in Friesz et al. [1989]; Janson [1991]; Papageorgiou [1990]; Ran et al. [1993, 1996]; Tong and Wong [2000]; Wie et al. [1990] adopt some form of link performance function (link travel time is expressed as a function of link inflow rate, outflow rate and/or vehicle accumulation, depending on the model). Akamatsu [2001]; Bernstein et al. [1993]; Huang and Lam [2002]; Kuwahara and Akamatsu [1997]; Kuwahara and Newell [1987]; Smith [1993] adopt a network of *point-queues* bottlenecks. As mentioned in §1.1.1, none of these representations is a sound approximation for the spatial propagation of congestion. Unfortunately, no model of equilibrium has convincingly incorporated traffic behavior based on the KW model yet. Recent attempts can be found in Lo [1999], which proposes a formulation for the ideal route-choice user equilibrium based on Daganzo's cell transmission model (no solution algorithm is proposed, though), and Kuwahara and Akamatsu [2001], which proposes an ad-hoc algorithm to solve the reactive equilibrium. However, no model yet combines commuter departure time choice with the *KW* model.

[7]Note that departure time choice is only consistent with predictive equilibrium since it is based on the assumption that arrivals times at the destination can be predicted.

1.1.3 Economic analysis: road pricing and investment

The ultimate objective of traffic prediction is controling the overall system so as to achieve an efficient use of the road infrastructure. This is specially important in the case of road networks since users tend to make decisions taking in account their individual trip costs, but disregarding the cost (i.e., delays) imposed onto other road users. This "selfish" behavior usually leads to more congestion than what would be optimal from a social point if everybody cooperated (what we called the system optimum).

Pigou [1920] first brought up this mismatch and fathered the concept of a congestion toll as a way improve road usage. Since then, economists have long studied the possibilities of road pricing. Economic modelling, though, has been largely based on the steady-state representation of traffic, as typified by the early works of Walters [1961] and Mohring and Harwitz [1962]. The basic paradigm is schematized in Figure 1.2. It uses the link performance functions of Figure 1.1 reinterpreted as an average travel cost vs. traffic demand (or number of trips made during the rush hour) curve, $AVC(q)$ in the figure. In agreement with the assumption of negative congestion externalities marginal cost is assumed to increase with demand as indicated by $MC(q)$.[8] At the same time, since traffic demand must logically depend on travel cost, a down-sloping demand curve $D(q)$ can be defined reflecting both the individual average and marginal value of using the road. The equilibrium traffic volume, q_{UE} is found at the crossing of the demand curve and the average cost curve. This equilibrium volume is higher than the *social optimal* usage, q_{SO}, found at the intersection of the marginal cost curve and the demand curve. To reestablish the *social*

[8] Cost are expressed as a function of traffic volume assuming an homogeneous monetary cost of time and considering vehicle operation costs independent of traffic level. As initially presented by Walters [1961], the cost curve also included a backward bending portion representing the congested branch of the fundamental diagram but, it is clear that for steady-state analysis this is not adequate.

optimum usage level, a toll equal to the difference in average and marginal cost at q_{SO} can be imposed. The optimal investment can also be analyzed since the travel cost curve depends on the road capacity level and the cost of providing capacity (i.e., the construction cost) can be reasonably estimated; see Mohring and Harwitz [1962] and Keeler and Small [1977]. A basic result is that congestion tolls will cover construction and maintenance costs over the long run in the presence of constant returns in road construction and maintenance. This single road pricing/investment model has been further enriched by considering users with different values of times, possible indivisibleness on the provision of capacity, etc.; see Hau [1998] for a review.

This basic static model can be also extended to network problems considering Wardrop's *user equilibrium* and *system optimum* principles [Wardrop, 1952]. Under the assumption that congestion on each link is a function of its volume, the system optimum can be achieved by imposing on each link an optimal toll of the same form as in the single road case [Beckmann et al., 1956]. Network road pricing modelling has been further enriched in different ways. For instance, models have been proposed to include different types of users (i.e., trucks, cars), to analyze second-best situations where only a reduced set of roads is priced, and to analyze situations where other modes of transport are also available. Reviews of the state-of-the-art on network road pricing models can be found in Lindsey and Verhoef [2000] and Arnott [2001]. The reliance on the classical steady-state model, however, has tilted the economic analysis towards considering trip quantity as the only factor of analysis, focusing excessively on pricing as the only solution (Arnott [2001] offers a good critique).

Dynamic settings have received less attention. The seminal work is Vickrey [1969] (already metioned above) where pricing and investment policies are evaluated under a time-varying congestion model represented by a single bottleneck queue. A

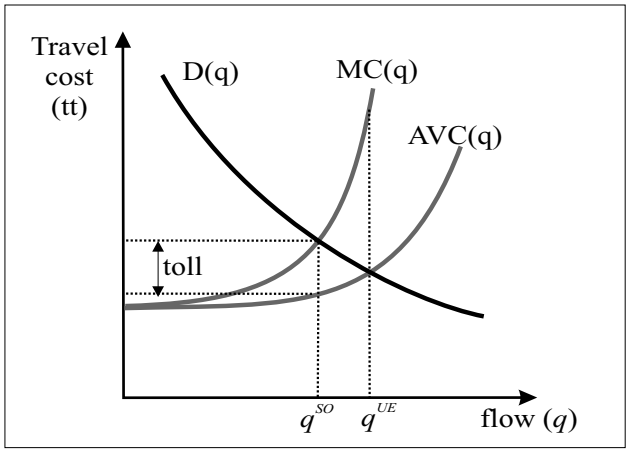

Figure 1.2. Congestion pricing model.

similar dynamic analysis is proposed in Henderson [1977] – later revisited and corrected in Chu [1992] – where a model of *local congestion* is used instead. Despite its highly idealized nature, Vickrey's bottleneck model unveiled some interesting insights particular to the dynamic case. For instance, unlike in the static case where congestion tolls always penalize road users, a time-dependent toll payment can leave all the commuters as well-off as before while turning wasted time in the queue into toll revenue. Vickrey's model has been extended to different scenarios (e.g., several parallel bottlenecks between a single OD pair, alternative modes), different demand assumptions (e.g., elastic mode-dependent demand, heterogeneous commuters) and different time-dependent toll policies (e.g., continuous toll, step toll); see Arnott et al. [1998] for a comprehensive review. More recent innovations include the analysis of mixed rationing/pricing schemes that prove to be *Pareto* improving for all road users [Daganzo and Garcia, 2000]. All these works are a substantial advance with respect to the traditional stationary analysis but their results are somehow limited because they do not explicitly consider spatial differences – congestion affects all commuters in a equal manner independently of their origins and destinations.

Kuwahara [1990] and Arnott et al. [1993a] extend Vickrey's analysis to consider two separate origins and a single destination, but the predictions in these works are based on a traffic model with point queues, which still limits the validity of the results. Wie and Tobin [1998] present a network traffic equilibrium model with dynamic pricing on each link, but again their model of traffic behavior does not consider properly queue propagation. Work that combines dynamic pricing with traffic models that account properly for queue propagation in space and time appears not to be done.

1.1.4 Urban location theory

A further step on the study of congestion recognizes the intimate link between land-use patterns and traffic congestion. Residential location choices generate a need for mobility which produces congestion and congestion, in turn, affects residential locations. Because individual location decisions affect the cost of living at other locations through increased congestion, an inefficient equilibrium distribution of population (e.g., with excessive sprawl) may arise if the congestion externality is not properly internalized through adequate pricing or land-use regulations.

The first models of urban location which explicitly incorporated congestion were proposed in Mills and de Ferranti [1971] and Solow [1973]. These work study the efficient provision of transportation infrastructure when congestion costs are properly internalized. The Mills/Solow framework abstracts from cumbersome network formulations and assumes a continuous mono-centric city where a distributed population travels to the *CBD* using a continuous network of radial roads. To represent traffic behavior, a model of *local congestion* is adopted where travel speed is only a function of the local traffic volume at each location (i.e., the number of vehicles crossing the

location) and independent of the conditions at any other location. Analytical and simulation results show that excessive sprawl and excessive land devoted to transportation happen if congestion is not priced adequately. A variety of works extend the Mills/Solow analysis to include different economic aspects. For example, Oron et al. [1973], Henderson [1975] and Arnott and MacKinnon [1978] primarily focus on inefficiencies in land use for housing under different market structures; Sullivan [1983a,b] explicitly consider labor markets; Wheaton [1998] compares land regulation and congestion pricing; Akai et al. [1998] analyzes equilibrium when transportation in provided by a private agent. Invariably, all adopt the local congestion model of traffic.

The results in all these works must be regarded with care for two main reasons, however. First, the *local congestion* model is highly unrealistic. As a result, the relationship between population sprawl and congestion are quite artificial. For example, it is easy to show that the level of congestion is dependent on the distance scale; i.e., if all the distances are doubled, the congestion cost are doubled. Therefore, results about optimal city size are irrelevant. Second, under the steady-state representation, the results are also clearly dependent on the order at which commuters are assumed to pass through each location, since this determines local traffic volumes and hence the congestion levels [Ross and Yinger, 2000]. For instance, the traditional Mills/Solow approach assumes that commuters join upstream users as they pass through their access location so that all commuters travel together and arrive to the common destination at the same time. On the other hand, Yinger [1993] assumes that commuters depart at the same time so that people at different location travel in different groups, or cohorts, and arrive at the destination ordered by distance to the destination, leading to an equilibrium location pattern different from those of Mills and Solow. To

make timing decisions endogenous, Ross and Yinger [2000] incorporates Henderson's model of dynamic equilibrium into the urban location problem, but it is concluded that no reasonable timing equilibrium can arise. This is yet another indication that local congestion is inadequate for the analysis.

In summary, finding an adequate way of incorporating a sensible model of flow propagation and commuter trip timing into the equilibrium model of urban location continues to be a challenge.

1.2 Dissertation overview

1.2.1 Scope

The literature review in the previous section shows that substantial effort has been devoted to the study of congestion in various fields. Although progress has been made in many areas - specially in the modelling of traffic dynamics - there are still substantial needs for improvement. The following four needs will be addressed in this thesis:

- **Models of morning commute need to better incorporate congestion propagation**. The available models of traffic equilibrium adopt unrealistic assumptions about traffic behavior. Either they ignore the spatial effect of queues or they assume that commuters are not spatially differentiated. Incorporating a realistic model into the network equilibrium analysis that can capture adequately the effect of physical queues will help addressing the aforementioned limitations.

- **Models of morning commute need to incorporate commuters trip tim-**

ing decisions. Insights on the effects of departure time choice on traffic equilibrium patterns are very limited. This is unfortunate since trip rescheduling is a likely commuter reaction to many policy measures. Since departure time decisions differ by location, it is necessary to extend the analysis of network problems to include departure time choice.

- **Models of morning commute need to focus on stylized scenarios**. Dynamic network equilibrium modelling has focused excessively on the development of algorithms rather than on the study of solutions; i.e., in the qualitative behavior of congested systems as a whole. The 'algorithmic' approach turns out to be rather inadequate to guide policy. Solutions can only be obtained through cumbersome computer-based simulation due to the complexity (and details) of network problems. From these solutions which are only particular to the scenario simulated, it is very difficult (if not impossible) to draw general qualitative insights about the behavior of congestion, which would be necessary to guide taxation and policy. Simplified scenarios that allow for the expression of the system behavior as a function of a few significant parameters and lead to analytical solutions may shed more light about the fundamental behavior of congestion.

- **Economic theory of urban location needs to revisited.** As shown in §1.1.3 and §1.1.4, the economic models of congestion have been largely based on steady-state and *local-congestion* assumptions. As a result, these models fail to adequately represent the true spatial behavior of congestion. Consistent relationships between the cost of congestion and the distribution of population need to be developed from realistic traffic models, such as the KW model, so that they can be further used for economic analysis.

The objective of our research is to relax as much as possible these four limitations. We develop a general model of the morning commute, which explicitly considers both realistic traffic behavior and commuters' departure time decisions in response to congestion. In addition, the dependence of departure time decisions and the distribution of population is explored. We seek to use the model to derive qualitative insights about the behavior of traffic in urban areas that will be valid in general. Therefore, the analysis will focus on selected geometries which include symmetries that allow for analytical solutions. Mono-centric cities, where commute is bound exclusively to a central business district (CBD) will be the main focus.

1.2.2 Main contributions

The main contributions of this research include:

- The development of the first model of traffic equilibrium which combines departure time choice and a realistic model of traffic flow. This model combines Vickrey's model of departure time choice with Newell's model of the KW theory.

- The development of an analytical procedure to solve the departure-time equilibrium for: (a) simple network with two-origins, (b) many-to-one tree networks.

- The theoretical analysis of the effects of ramp metering and capacity expansion when departure time is an issue. This analysis reveals unexpected situations where ramp-metering can be beneficial, and others where the provision of more freeway capacity or storage can be counterproductive.

- The study of the relationship between congestion cost and population distribution for mono-centric cities. This study is based in both discrete (network-based) and continuous models.

- The development of closed-form expressions that link congestion costs to location and spatial population distribution. These formulae are structural relationships since they make endogenous commuters timing decisions and traffic dynamics; hence, they can be generally applied to study other more general urban problems.

1.2.3 Organization

The thesis is organized in a series of self-contained chapters. Chapter 2 presents a first model that explicitly considers the most important determinants of congestion behavior during the morning commute: different commuter origins, merge interactions and queue spillovers. We examine the simplest possible network (2 origins and one destination) exhibiting the three important features. This model can be used as a building block for the analysis of more complex, single-destination networks with departure-time choice. Chapter 3 extends the analysis to the case of a long homogeneous freeway. This model is relevant since a long homogeneous freeway is the logical unit of analysis for mono-centric cities with ring-radial street networks. We develop an exact analytical procedure that can be used to model morning commute traffic evolution in long corridors. The analysis of congestion in monocentric cities is presented in chapter 4. Both a discrete and continuous formulation are investigated. General closed-form solutions are proposed that allow the quantification of commuting costs as a function of location and population distribution. The focus of chapters 2, 3 and 4 is on concepts, qualitative insights and policy analysis, rather than in methodologies. For that purpose, we adopt some simplifying assumptions that allow direct analytical treatment; e.g., that the networks are homogenous and commuters have the same desired arrival time to the destination. Chapter 5 - which can be

considered a technical addendum to the previous chapters - extends the analysis to more general instances where the network is nonhomogeneous and commuters have different desired arrival times. This chapter provides additional insight and discusses the difficulties one encounters in the design of solution algorithms for the morning commute problem over general networks. Finally, chapter 6 presents some conclusions and discusses possible extensions of the work in this thesis.

Chapter 2

A Simple Network Model

Vickrey [1969] describes the first traffic model where commuters can adapt their departure time to avoid periods of high congestion. The model is very simple – a single bottleneck with a fixed number of commuters – but also very revealing of possible policy actions for congestion reduction. Because of its appeal and simplicity, Vickrey's model has been extensively analyzed under different demand assumptions [Arnott et al., 1993b; Daganzo, 1985; Hendrickson and Kocur, 1981; Newell, 1987; Smith, 1984] and has also been adopted to analyze various toll policies [Arnott et al., 1990; Daganzo and Garcia, 2000; Laih, 1994]. The model, however, only applies to cases where congestion is concentrated at a single location, affecting all commuters equally. These conditions are violated when the access network is itself congested. For example, freeway queues caused by bottlenecks often spill over long distances imposing different penalties on its access points. Obviously, network effects should be investigated.

This chapter introduces a network model that integrates Vickrey's theory with a realistic traffic flow model [Newell, 1993] and a reasonable merging mechanism [Daganzo, 1995a]. Our ultimate goal is the qualitative understanding of the relationship

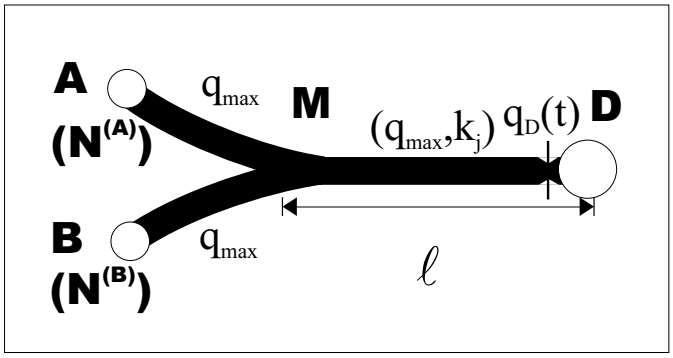

Figure 2.1. Homogeneous 2-origin network.

among congestion, departure time choice and the spatial distribution of population for the morning commute, recognizing the networks are congested and have different origins. We consider the simplest network with all these relevant characteristics. It consists of two origins, one destination and two links merging into a third, as shown in Figure 2.1.

The chapter is structured as follows. Section 2.1 introduces relevant background and discusses the single bottleneck (Vickrey) model. Section 2.2 presents the equilibrium and the traffic model for the two origin network. Section 2.3 presents the results. Section 2.4 compares the solutions with those obtained under *point-queue* assumptions. Finally, section 2.5 discusses policy implications and relates them to earlier work.

2.1 The single bottleneck model

It is commonly assumed that traffic conditions during the morning commute are similar day-after-day. Commuters, aware of these, choose their departure time to minimize their individual trip cost, which consist of a trip-time component and a

schedule penalty. The latter is associated with the actual arrival time at the destination relative to a *preferred* arrival time. In the case where the only traffic restriction is a single bottleneck of capacity q_D with no delay elsewhere, it is customary to express commuter decisions as a function of the *preferred passage time* through the bottleneck or *deadline*. If w_t is the deadline for the commuter that passes the bottleneck at time t and we express costs in units of trip time, then the trip cost for that commuter is

$$c = \tau + p(t - w_t), \tag{2.1}$$

where τ is the trip time and $p(\cdot)$ is a schedule penalty function such that $p(\cdot) \geq 0$ and $p(0) = 0$. It will be assumed here that $p(\cdot)$ is piecewise linear and V-shaped, where e and L are the positive conversion rates for *earliness* and *lateness* into trip time; i.e.,

$$p(s) = \begin{cases} -es & \text{if } s < 0 \\ Ls & \text{if } s \geq 0. \end{cases} \tag{2.2}$$

Normally, earliness is preferable to both queuing and lateness, i.e., $e < L$, $e < 1$ [Small, 1982]. The objective is then determining an equilibrium schedule of departures from a single origin such that no commuter/vehicle would have an incentive to change its departure time given the queues that resulted from the equilibrium. The model also applies to multiple origins if all access routes to the bottleneck are uncongested and pass through a common point, O; i.e., point O can be modelled as the single origin.

The solution can be represented by means of continuous cumulative plots, assuming that the number of commuters is so large that vehicles can be treated as a continuous variable; see Figure 2.2. $W(t)$ expresses the cumulative number of commuters wishing to pass the bottleneck by time t, and it will be called the *deadline*

curve. It will be assumed that $W(t)$ is S-shaped, with slope greater than the capacity q_D during some interval so that a queue must necessarily develop. $W(t)$ is a step function if all the commuters have the same deadline, as shown in Figure 2.2a. Then, the objective is finding an equilibrium curve of cumulative arrivals at the common point O, $A_O(t)$ – or equivalently the curve of cumulative virtual arrivals at the bottleneck, $A(t) = A_O(t - t_{OD})$ where t_{OD} is the fixed uncongested trip time from O to the bottleneck location, D.[1] According to standard queuing analysis, the curve of cumulative *departures* from the bottleneck, $D(t)$, is the highest curve with slope less than or equal to q_D such that $D(t) \leq A(t)$.[2] Under a FIFO (*first-in-first-out*) queue, the delay τ for any given vehicle number is the horizontal distance between curves A and D. Likewise, the scheduled delay s is given by the horizontal distance between D and W if vehicles depart from the bottleneck in the order of their deadlines. It is known that if the penalty function $p(\cdot)$ is convex and common to all commuters, the solution exists [Smith, 1984] and is unique [Daganzo, 1985]. Furthermore, in the equilibrium solution, vehicles depart from the bottleneck in the order of their deadlines. An example of such equilibrium is represented in Figure 2.2 both for the case when commuters have a common deadline (Figure 2.2a) and when they do not (Figure 2.2b). Both solutions exhibit a unique queuing episode with two clearly differentiated phases. In the first phase, commuters depart from the bottleneck earlier than desired and queuing delay increases with vehicle number at a rate that precisely compensates for the reduction in earliness. Therefore, the slope of $A(t)$ is given by $q_D/(1 - e)$. In the second phase, commuters depart from the bottleneck later than

[1] A *vehicle virtual arrival time* to D is the time at which the vehicle would have passed D if it had travel unhindered from O to D.

[2] In queuing lingo, the terms *arrivals* and *departures* refer to the bottleneck. Therefore, they have the reverse meaning assigned to them in the economics literature where *arrivals to the bottleneck* correspond to the *departures from the origin* and vice-versa.

desired and queuing time declines with vehicle number to compensate for increasing lateness penalties. As a result, the slope of $A(t)$ is also given and equal to $q_D/(1+L)$. Note that the vehicle arriving on time experiences the highest delay as given by the length of segment **AO**, $|\mathbf{AO}|$, in Figure 2.2. In the single deadline case of Figure 2.2a, $|\mathbf{AO}|$ is the common cost suffered by all commuters.

If t_s and t_f are the times when the queue starts and vanishes, equilibrium requires $|\mathbf{AO}| = Lt_f = -et_s$. Furthermore, if we use N to denote the number of commuters who queue, then $N = q_D(t_f - t_s)$ since all these commuters depart when the bottleneck is at capacity. In the single deadline case, N is known (i.e., all the commuters suffer delay), therefore these three equations define the three remaining unknowns: $|\mathbf{AO}|$, t_s and t_f. Since the slopes of $A(t)$ above and below **AO** are given, it follows that there is only one possible geometry for the equilibrium curves. Figure 2.2a shows that the number of commuters departing early at equilibrium is $NL/(e+L)$ and the number departing late is $Ne/(e+L)$. One can also see that the common cost is $|\mathbf{AO}| = NeL/(e+L)$. Finally note that the equilibrium delay for a commuter departing at time t, $\tau(t)$, is

$$\tau(t) = \begin{cases} |\mathbf{AO}| + et = e(t - t_s) & \text{if } t < 0 \\ |\mathbf{AO}| - Lt = e(t_f - t) & \text{if } t \geq 0 \end{cases}, \tag{2.3}$$

which precisely balances the schedule penalty as required.

Consideration shows that a similar geometric pattern is an equilibrium for any S-shaped deadline curve; see Figure 2.2b. The main difference is that in this case not all the commuters queue; therefore, one also needs to find N.

Most of the existing literature deals with fixed-capacity bottlenecks, but the analysis can be extended to variable capacities, $q_D(t)$. Then, t_s, t_f and $A(t)$ can be

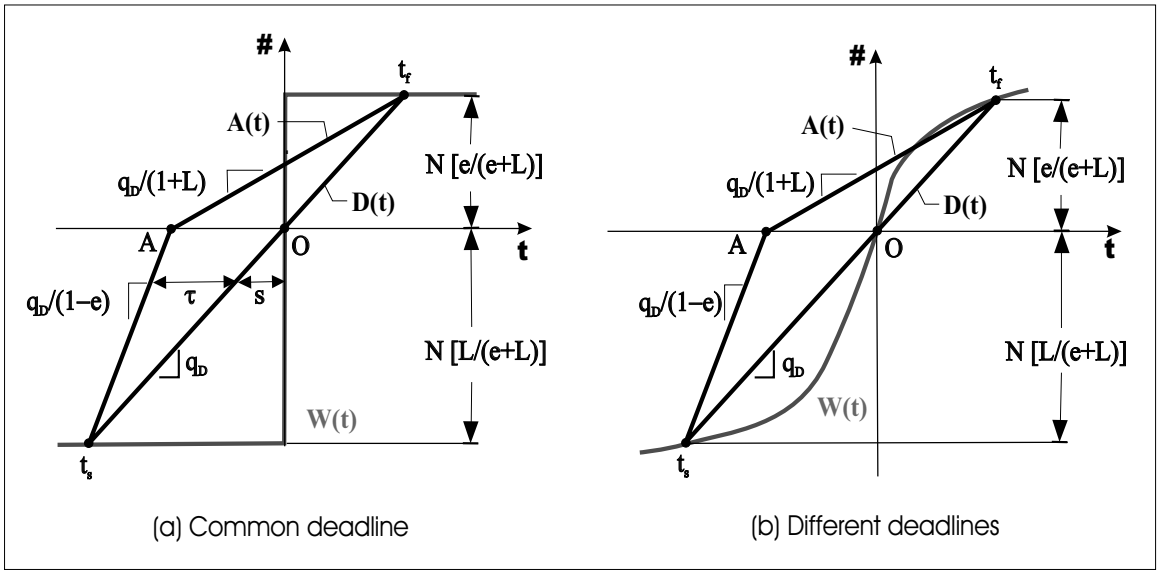

Figure 2.2. Single bottleneck equilibrium solution (fixed capacity).

determined as before since (2.3) continues to hold. This means that in any equilibrium, such as that shown in Figure 2.3a, the horizontal separation between A and D at $\# = D(t)$ continues to be given by (2.3). Therefore, if the equilibrium diagram is rescaled vertically by means of the transformation $\overline{\#} = D^{-1}(t)$ which makes the departure rate equal to 1 at all times, i.e., $D^{-1}(D(t)) = t$, then we recover Figure 2.2a. This is shown in Figure 2.3b. The re-scaled arrival curve, $T(t) = D^{-1}(A(t))$, now returns the departure time t_d (on the vertical axis) as a function of the arrival time t_a (on the horizontal axis). We shall refer to $T(t)$ as the *arrival-departure schedule* curve (or A/D curve) to differentiate it from the actual equilibrium arrival curve, $A(t)$. The invariance of the rescaled diagram with respect to $q_D(t)$ will become useful later.

It should be remembered that the single bottleneck model does not apply if delays experienced by vehicles entering the network at different locations are different, as is normally the case for freeway networks. Unfortunately, no existing model addresses

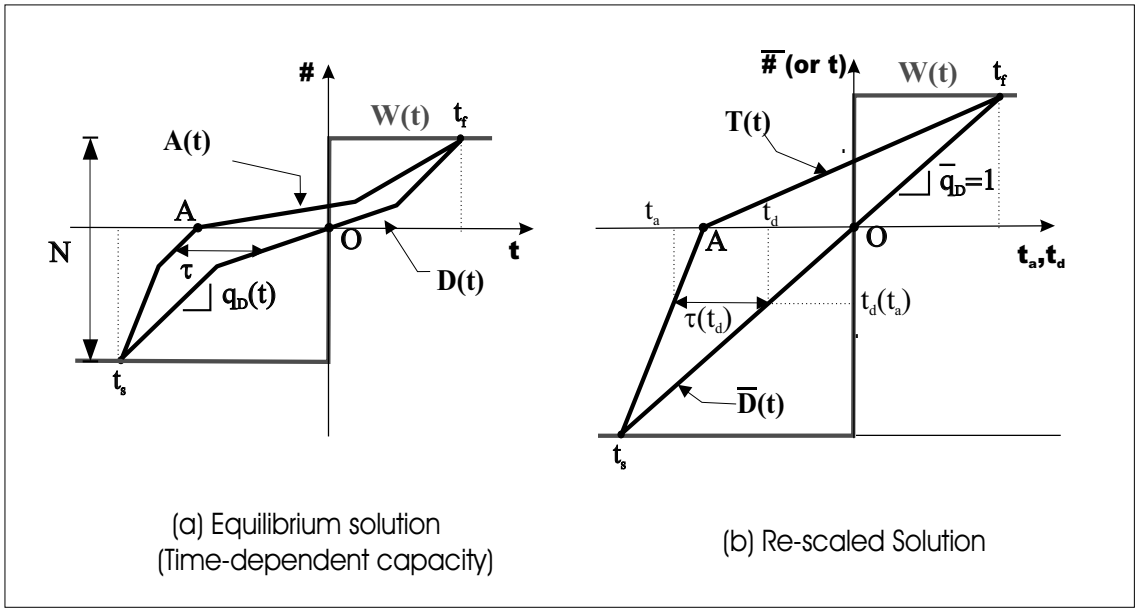

Figure 2.3. Single bottleneck equilibrium solution (time-dependent capacity).

the three key effects required to model a simple freeway: multiple origins, merging interactions and queue spillovers. The next section describes a first step in this direction.

2.2 Two-origin departure-time equilibrium and realistic traffic behavior

2.2.1 Problem formulation

We shall consider here the simplest network exhibiting all three effects; see Figure 2.1. On this network, $N^{(A)}$ and $N^{(B)}$ commuters travel everyday from origins A and B to a common destination D. The routes from these origins merge at an intermediate location, M, and share a final link MD of length ℓ. A bottleneck of (possibly

time-dependent) capacity $q_D(t)$ may exist just upstream of D and queues may form on the common link and spill over the merge.[3] For simplicity, we assume that: (1) the network is homogeneous (i.e. its three links have the same characteristics), (2) all commuters have the same deadline and penalty function (i.e., commuters are only distinguishable by their origin). Generalizations for networks with non-homogeneous links and different deadlines are discussed in chapter 5.[4]

We express our solution in terms of *origin-specific* cumulative inflows (or cumulative departures from each origin) and cumulative outflows from point D (or arrivals to the destination). We can ignore free-flow trip times in our analysis, since those are fixed for each origin and independent of the time of arrival. In this case, the solution is defined by the curves $\{A^{(r)}(t), r = A, B\}$, which represent the *virtual cumulative arrivals* at point D for each origin – instead of the actual cumulative departure curves from origins A and B – and $\{D_D^{(r)}(t), r = A, B\}$, the cumulative *departures* from D. (From now on, superscripts identify the origin to which the variable or function refers, while subscripts refer to the physical location over which the variable or function is defined. Furthermore, $a^{(r)}$ and $d_D^{(r)}$ represent the time-derivatives or flows respectively; e.g, $d_D^{(r)}(t)$ is the flow at D of commuters from origin r.) Delays – instead of actual travel times - are given by the horizontal separation between $A^{(r)}$ and $D_D^{(r)}$; i.e., $\tau^{(r)} = t - {A^{(r)}}^{-1}(D_D^{(r)}(t))$. The actual departure curves from the origins can be obtained by shifting the virtual curves back in time by the origin-specific free-flow trip times; i.e., $A_r^{(r)}(t) = A^{(r)}(t + \ell_{rD}/v_f)$ where ℓ_{rD} is the total distance from origin r to point D and v_f the free-flow speed. Actual trip times are given by $\tau_{rD}^{(r)} = \tau^{(r)} + \ell_{rD}/v_f$.

A possible assignment pattern (not necessarily in equilibrium) is presented in

[3] The flow restriction could be due to a variable inflow from another ramp (not depicted in Figure 2.1) very close to D.

[4] A summary of the notation used in this chapter and throughout the dissertation can be found in appendix B.

Figure 2.4. The two origin-specific diagrams of Figure 2.4a can be conveniently superimposed, by adding vehicle numbers, to analyze the traffic behavior on link MD as shown in Figure 2.4b. The curves D_M and D_D represent the actual departure curves from M and D, respectively. The proportion of departures by origin at time t is defined as $\alpha_D^{(r)}$ (i.e., $d_D^{(r)}(t) = \alpha_D^{(r)}(t) d_D(t)$). Under FIFO conditions, the delays experienced in link MD, τ_{MD} – given by the horizontal distance between D_M and D_D – must be equal for both origins. Furthermore, the proportion of vehicles departing from D must be the same when the vehicles passed M, i.e., $\alpha_M^{(r)}(t - \tau_{MD}(t)) = \alpha_D^{(r)}(t)$. Finally, it is convenient to consider the re-scaled *origin-specific A/D* curves $T^{(A)}$ and $T^{(B)}$ constructed as explained in §2.1 since these curves allow recovering the actual experienced delays. Total delays are given by the horizontal distance between $T^{(r)}$ and D_D; delays in each approach upstream of the merge by the horizontal distance between $T^{(r)}$ and D_M. We will make extensive use of this construction when analyzing the solutions.

2.2.2 Traffic dynamics

The equilibrium solutions must then be consistent with both link and node dynamics. Basically, the two phenomena that affect the traffic solution are the queuing behavior on link MD and the merge interactions. Traffic in link MD is modelled as in the simplified *kinematic wave* (KW) theory proposed in Newell [1993]. According to the theory, traffic obeys a triangular fundamental relationship linking flow q with density k defined by three parameters: a fixed free-flow speed (v_f), a maximum flow or capacity ($q_{\max}$) and a jam density (k_j);[5] see Figure 2.5a. Newell shows that the *delay-based* traffic problem can be solved as a standard problem with the modified

[5] Jointly they define a wave speed w which represents the unique speed at which flow disturbances propagate upstream within a moving queue.

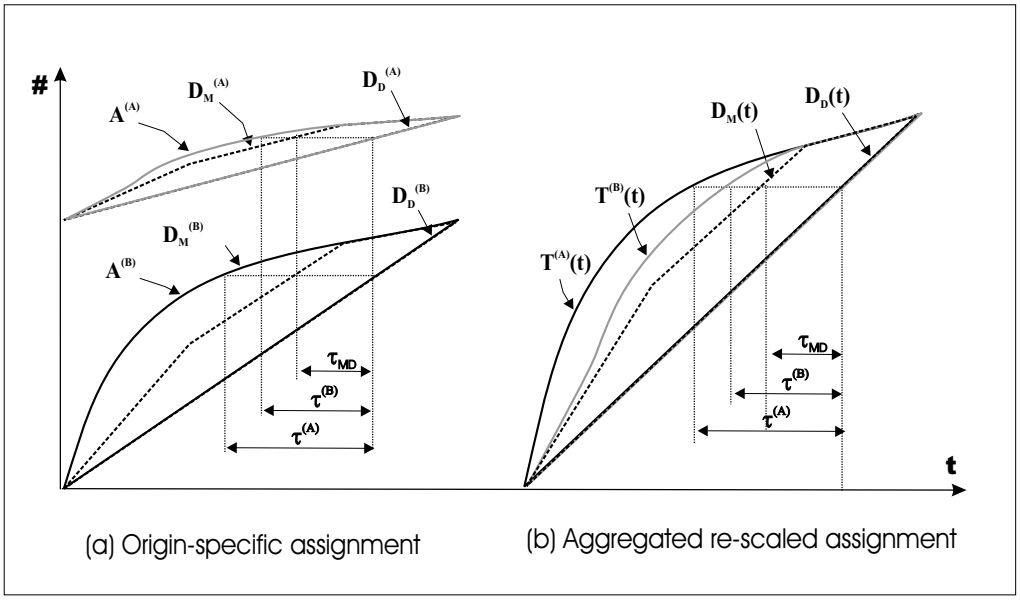

Figure 2.4. Traffic assignment with 2 origins. Cumulative plot representation.

fundamental diagram of Figure 2.5b, which has the same $q_{\max}$ and k_j but $v_f = \infty$. The traffic model is completed by defining how vehicles interact at the merge. We will use the rules in Daganzo [1995a]. These are depicted in Figure 2.5c and explained later.

Physical queues dynamics: The delays in link MD must be predicted since we must guarantee that commuters from different origins passing M at the same time incur the same delay on link MD (i.e., queues are FIFO). Physical queue are relevant in link MD since the common delays on the link depend on the queue spilling over the merge section or not. For links AM and BM, physical queues are not an issue because the delays suffered in these links are always common to all the commuters from the same origin.

According to Newell, a capacity curve at M, D_M^D, is defined from the departure curve at D, D_D, by the shift,

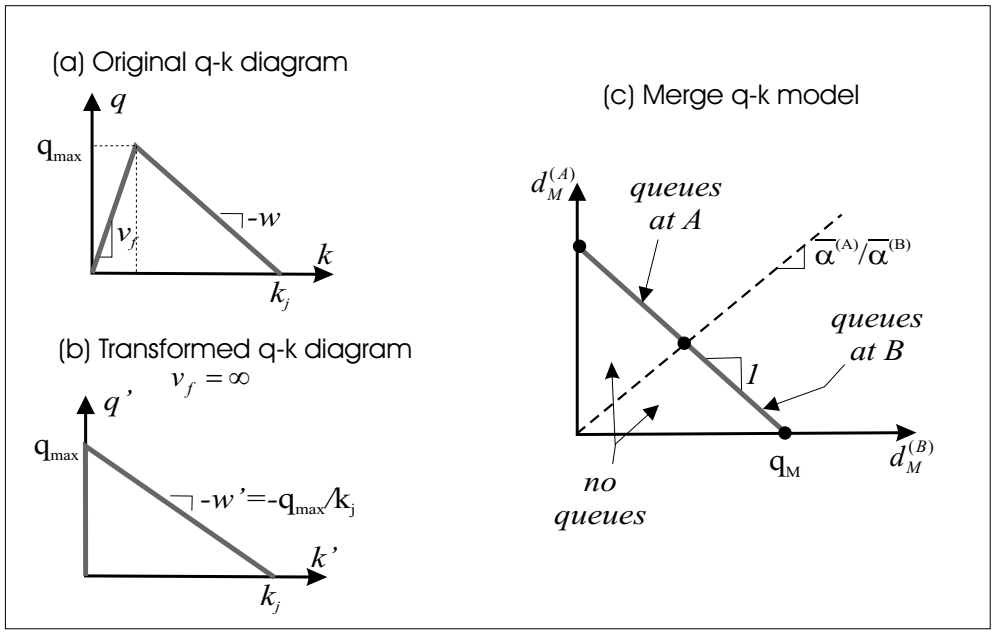

Figure 2.5. Traffic flow model and merge model [Newell, 1993].

$$D_M^D(t) = D_D(t - \frac{k_j}{q_{\max}}\ell) + k_j\ell \tag{2.4}$$

This capacity curve tracks the effects of the backward moving queue on the entrance of link MD and sets an upper bound to the cumulative number of commuters that can pass M by time t; see Figure 2.6. The actual cumulative curve of vehicles passing through M, $D_M(t)$, is the lower envelope of D_M^D and the cumulative number of commuters who would have passed M in the absence of a queue, which is determined by upstream demand. In our case, the upstream behavior depends on the merge behavior.

Merge interactions: The flows from the two approaches must share the downstream capacity according to some pre-specified merging rules. Daganzo [1995a, 1996] proposes that the upstream flows from each merging approach – $d_M^{(A)}, d_M^{(B)}$ – must be a

function of the capacity of the upstream approaches (i.e., $q_{\max}$), the available capacity downstream (q_M) and some *approach-specific* priority ratio – $\tilde{\alpha}^{(A)}, \tilde{\alpha}^{(B)}$ – where $\tilde{\alpha}^{(A)} + \tilde{\alpha}^{(B)} = 1$. For the case of interest here where the (time-dependent) downstream capacity is given by $q_M(t) = d_M^D(t) < q_{\max}$, the upstream approach capacities do not play a role and the rules reduce to the following two:

(1) during periods when there are no queues upstream of M, arrival flows equal discharge flows and $d_M = d_M^{(A)} + d_M^{(B)} \leq q_M(t)$;

(2) when there is a queue on approach r, then $d_M = q_M(t)$ and the departure ratio $\alpha_M^{(r)} \equiv d_M^{(r)}/d_M \geq \tilde{\alpha}^{(r)}$.

It follows from (2) that when there is a queue in both approaches, then $d_M^{(A)}/d_M^{(B)} = \tilde{\alpha}^{(A)}/\tilde{\alpha}^{(B)}$. These rules are illustrated in Figure 2.5c. A more detailed description of the dynamics of the merge section for more general cases can be found in Daganzo [1995a, 1996] and in chapter 3, §3.2.

Delay-based representation: In our case, it is convenient to express all the traffic feasibility conditions in terms of the some candidate equilibrium departure curves from D – $D_D^{(A)}$ and $D_D^{(B)}$ – and the origin-specific equilibrium delays – $\tau^{(A)}$ and $\tau^{(B)}$ – indexed by time of departure. As we shall show later in §2.2.3, equilibrium conditions are easily expressed as a function of these functions.

To express behavior in link MD as a function of delays, first note that D_D will be such that

$$d_D(t) = q_D(t) \text{ if } \max\{\tau^{(A)}(t), \tau^{(B)}(t)\} > 0 \tag{2.5}$$

Furthermore, note that $D_M^D(t) - D_D(t)$ is an upper bound for the length of the queue at MD at time t. Hence, the horizontal separation between D_M^D and D_D, $\nu_{MD}(t)$,

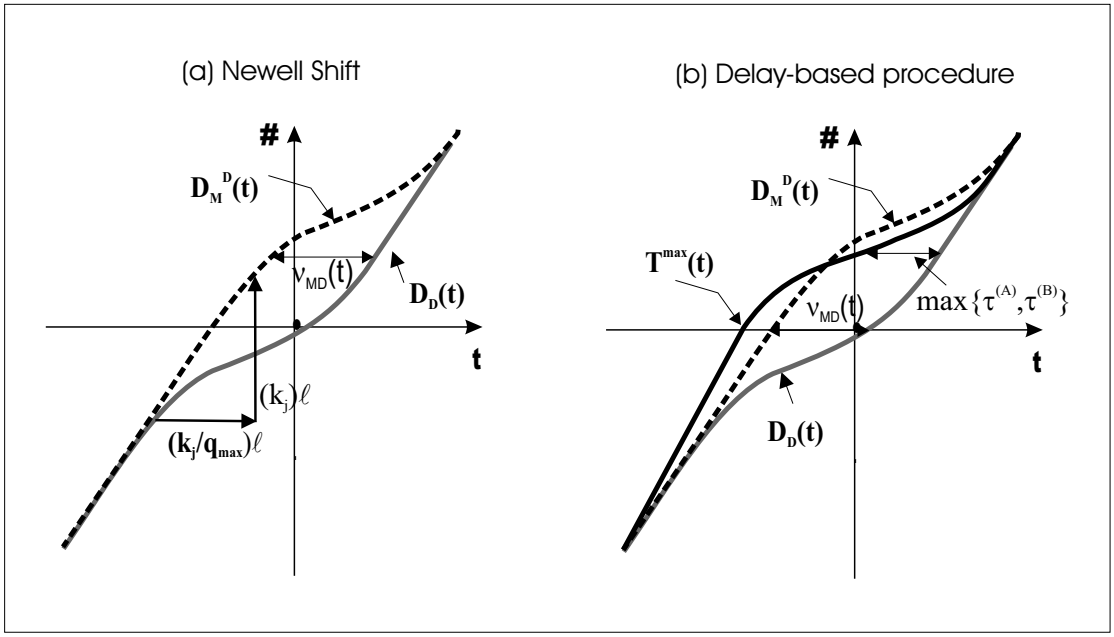

Figure 2.6. Newell's KW procedure.

is also an upper bound for the delay on link MD for any given departure time (we call $\nu_{MD}(t)$ the maximum delays or M-delays). We can compare the delays $\tau^{(A)}$ and $\tau^{(B)}$ with ν_{MD} in an attempt to infer the actual delays τ_{MD} and hence, the actual departure curve from M, D_M. Since $0 \leq \tau_{MD} \leq \nu_{MD}$ and $0 \leq \tau_{MD} \leq \max\{\tau^{(A)}, \tau^{(B)}\}$ (because of the FIFO discipline), then three main situations cases can arise, as given by

$$d_M(t_M) = \begin{cases} d_M^D & \text{if } \nu_{MD} = \tau_{MD} \leq \max\{\tau^{(A)}, \tau^{(B)}\} \\ q_{\max} & \text{if } \tau_{MD} < \max\{\tau^{(A)}, \tau^{(B)}\} < \nu_{MD} \\ min\{a^{(A)} + a^{(B)}, q_{\max}\} & \text{if } \tau_{MD} = \max\{\tau^{(A)}, \tau^{(B)}\} < \nu_{MD} \end{cases} \tag{2.6}$$

The first equality corresponds to the case with queues spillovers at M; the second, to the case with no queue spillover at M but queues at the upstream approaches; the

third, to the case with queues only downstream of M. Based on (2.6), the actual curve D_M and the delays τ_{MD} can be defined as function of $\max\{\tau^{(A)}, \tau^{(B)}\}$ in the following graphical manner (see Figure 2.6): draw the curve $T^{\max}$ such that the horizontal distance between $T^{\max}$ and D_D for each t corresponds to $\max\{\tau^{(A)}(t), \tau^{(B)}(t)\}$ (i.e., $T^{\max}$ is the A/D curve for the maximum delays); then, obtain D_M as the lower envelope of D_M^D and the higher curve underneath $T^{\max}$ with slope $\leq q_{\max}$. The delays τ_{MD} are given by the horizontal distance between D_D and D_M for each departure time t, i.e., $\tau_{MD}(t) = t - D_M^{-1}(D_D(t))$.

The *merging rules* must also be expressed in terms of our candidate departure and delay curves. When $\tau^{(r)}(t) > \tau_{MD}(t)$, vehicles from approach r departing from D at t experience delay upstream of the merge. Hence, in view of merge condition 2, they must enter the merge in a proportion $\alpha_M^{(r)} \geq \tilde{\alpha}^{(r)}$. Since this proportion must be preserved at D, we can write:

$$\tau^{(r)}(t) > \tau_{MD}(t) \Rightarrow \alpha_D^{(r)} \geq \tilde{\alpha}^{(r)} \tag{2.7}$$

$$\tau^{(A)}(t) > \tau_{MD}(t), \tau^{(B)}(t) > \tau_{MD}(t) \Rightarrow (\alpha_D^{(A)}, \alpha_D^{(B)}) = (\tilde{\alpha}^{(A)}, \tilde{\alpha}^{(B)}). \tag{2.8}$$

Condition (2.7) applies when the merge is at capacity with one congested approach and (2.8) with two. When the merge is non-congested, any proportion of flows is allowed. Finally, if $\tau^{(r)}(t) < \tau_{MD}(t)$ the solution is only feasible if $d_D^{(r)}(t) = 0$.

2.2.3 Equilibrium conditions

Departure time-equilibrium requires that the trip cost for commuters in the same origin r must be equal for any chosen arrival time and equal or larger for any other non-chosen times. The traffic-equilibrium problem could be solved by considering the

arrival curves $\{A^{(r)}, r = A, B\}$ as the unknowns since these define through the traffic model unique $\{D_D^{(r)}, r = A, B\}$ and $\{\tau^{(r)}, r = A, B\}$. One would need to iterate on the $A^{(r)}$'s until finding costs that satisfy the equilibrium criterion.

The equilibrium solution, however, can be more easily obtained in one takes in account some properties of the solution.

Property UE(1) (*or Parametric representation of equilibrium).* The equilibrium delays (or trip times) for each origin, $\tau^{(r)}$, are uniquely determined by the time of arrival of the first commuter from the origin suffering any delay, $t_s^{(r)}$, and so is the equilibrium cost, $\tilde{C}^{(r)}$.

Proof. Recall that in the single origin the equilibrium delay $\tau(t)$ was affected by $D(t)$ through t_s and t_f. Obviously, the same simple principle applies now but with origin-specific $t_s^{(r)}$ and $t_f^{(r)}$. From equation (2.1), the equilibrium cost for the commuter arriving at $t_s^{(r)}$ is given by $\tilde{C}^{(r)} = p(t_s^{(r)})$ and thus,

$$\tau^{(r)}(t|t_s^{(r)}) = \tilde{C}^{(r)} - p(t) = p(t_s^{(r)}) - p(t). \blacksquare \tag{2.9}$$

Property UE(1) suggests the following solution approach. Choose initially a set $\{t_s^{(r)}, r = A, B\}$, which defines unique *origin-specific* equilibrium delays (or travel times), $\{\tau^{(r)}$,r=A,B$\}$. Then, find the traffic-feasible arrival and departure processes $\{A^{(r)}, D_D^{(r)}, r = A, B\}$ which are consistent with the $\{t_s^{(r)}, r = A, B\}$ and $\{\tau^{(r)}$,r=A,B$\}$. We can use the delay-based traffic formulation in §2.2.2 to do this. This traffic assignment $\{A^{(r)}, D_D^{(r)}, \tau^{(r)}\}$ will be an equilibrium but the total outflows $\{D_D^{(r)}(t_f^{(r)}), r = A, B\}$ may not match the populations $N^{(r)}, r = A, B$. Hence, we would have to change the $t_s^{(r)}$'s until $\{D_D^{(r)}(t_f^{(r)}), r = A, B\}$ are equal to $\{N^{(r)}, r = A, B\}$.

An immediate corollary of property UE(1) is

Property UE(2) (*or sequential ordering of delays*). Given an ordering by origin of initial times $t_s^{(r)}$ (or equivalently, of costs $\tilde{C}^{(r)}$), the equilibrium trip times for any other departure time follow this same order, i.e., $\forall r, s\ \tilde{C}^{(r)} > \tilde{C}^{(s)} \Longleftrightarrow \tau^{(r)}(t) > \tau^{(s)}$.

Property UE(2) states that the maximum delays experienced by any commuter for a given departure time, $\max\{\tau^{(A)}, \tau^{(B)}\}$, necessarily coincide with the equilibrium delays for one of the origins. This suggests that the solution procedure can be streamlined even more by considering only equilibrium solutions where the full capacity available at D is utilized during the interval $\Pi = [t_s, t_f]$ given by the single origin bottleneck solution with total population $N^{(A)} + N^{(B)}$, i.e., the combined departure curve at D, D_D, coincides with that of a single origin problem. This is intuitive since our network model allows un-delayed travel from both origins when section D is under capacity (therefore, commuters independently of the origin they come from would have an incentive to use the full capacity at the preferred times).

2.3 Equilibrium analysis

We consider the single deadline problem and analyze it in two phases: (i) cases with no bottleneck restriction at D ($q_D(t) \equiv q_D = q_{\max}$) where queues cannot form on link MD and merging effects dominate, and (ii) cases with time-dependent flow restrictions at D ($q_D(t) \leq q_{\max}$) where queue spillovers can affect performance.

2.3.1 No downstream restrictions (Merging effect)

When no restrictions exist downstream of merge M, no delays can arise beyond it. Therefore, we can ignore link MD and treat M as if it was the destination using $D_M(t) \equiv D_D(t)$, $\alpha_M^{(r)}(t) \equiv \alpha_D^{(r)}(t)$. In essence, the system is modelled as a pair

of single-origin bottlenecks with departure rates coupled by the merging rule. As mentioned in §2.2.3, it is logical to consider equilibrium solutions in which the aggregated departure curve D_M is given by the single bottleneck solution with capacity q_D and total population $N^{(A)} + N^{(B)}$, i.e., where the merge is saturated only during the preferred interval $\Pi = [t_s, t_f]$ of Figure 2.2a.

Since D_M is given, the capacity shares $\{\alpha_M^{(A)}(t), \alpha_M^{(B)}(t)\}$ need to be found. If we further consider that only one queuing episode can occur on each approach, the solution is as shown in Figure 2.7a. The figure displays the arrival and departure pattern on the two approaches separately. Commuters from one origin (B in the figure) flow through the bottleneck during an interval when the other approach is queued. Therefore, from (2.8), they use a fixed share of the capacity $\alpha_M^{(B)}(t) = \tilde{\alpha}^{(B)}$. The solution for B-users is a single bottleneck equilibrium with population $N^{(B)}$ and capacity $\tilde{\alpha}^{(B)} q_D$; see the bottom part of Figure 2.7a, curves $D_M^{(B)}$ and $A^{(B)}$. Commuters from A flow at full capacity q_D when the approach B is not active and at a reduced capacity $\tilde{\alpha}^{(A)} q_D$ otherwise. The solution for A-users is also a single bottleneck equilibrium, albeit with time-dependent capacity; see the top part of Figure 2.7a, curves $D_M^{(A)}$ and $A^{(A)}$.

The two diagrams of Figure 2.7a can be re-scaled and superimposed following the procedure explained in §2.2.2 to show the *A/D curves* for both origins and the common departure curves on a single diagram, see Figure 2.7b. The quantities in parenthesis following each colon are $\{\alpha_M^{(A)}(t), \alpha_M^{(B)}(t)\}$. The curves on the figure must be similar, as shown, since commuters share a common deadline and penalty function. Commuters from A experience the same commuting cost as if everybody had the same origin, since **AO** is equal in length to the corresponding segment of a single bottleneck solution with population $N^{(A)} + N^{(B)}$. Commuters from B, however, experience a

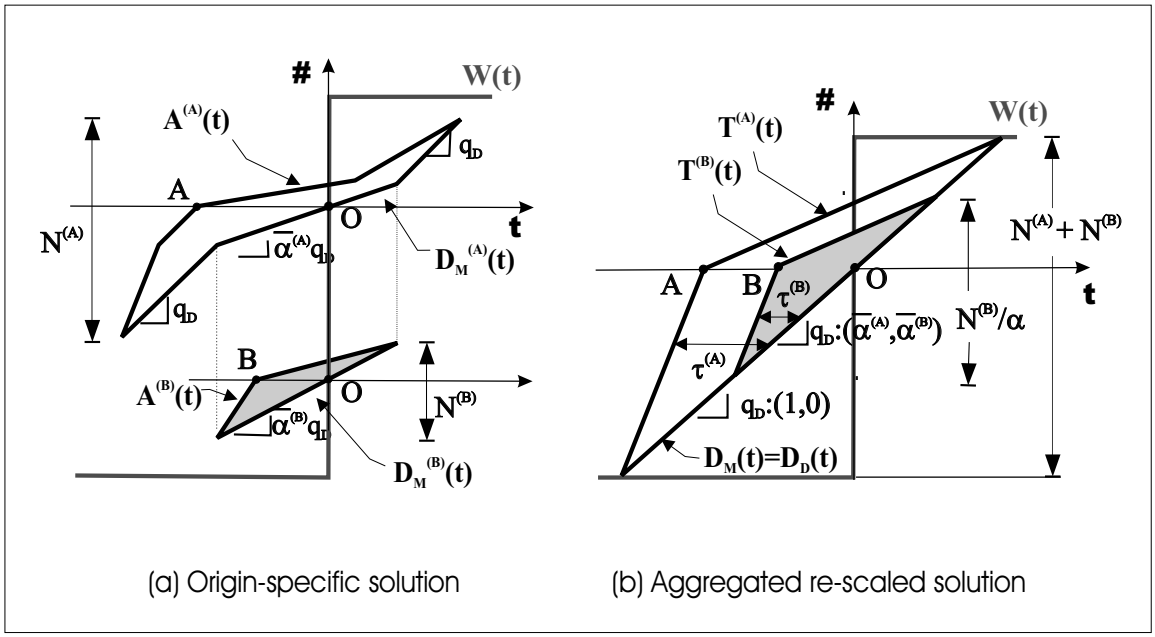

Figure 2.7. Equilibrium solution, no queues in link MD.

reduced cost, $|\mathbf{BO}|$. From the figure, it is clear that B-commuters experience less cost if $N^{(B)}/\tilde{\alpha}^{(B)} < N^{(A)}/\tilde{\alpha}^{(A)}$. The reverse is true if $N^{(B)}/\tilde{\alpha}^{(B)} > N^{(A)}/\tilde{\alpha}^{(A)}$. The worst case arises if $N^{(B)}/\tilde{\alpha}^{(B)} = N^{(A)}/\tilde{\alpha}^{(A)}$ when all commuters experience the highest cost. The best case arises if $\tilde{\alpha}^{(A)} = 0$ or 1 (complete priority) since one of the origins experiences the least possible cost.

These results have an economic interpretation. Since the capacity of M is a scarce resource, commuters impose onto each other an external cost (delay) as they jockey for their preferred departure times. In the single bottleneck scenario, everybody is affected equally by the actions of the others and the result of this game is a symmetric equilibrium. A saturated merge, however, allocates its capacity in fixed shares – $\tilde{\alpha}^{(A)}$ and $\tilde{\alpha}^{(B)}$ – to the two approaches, which insulates B-drivers from actions of A-drivers and vice-versa. This allows one part of the population to reduce its cost by traveling at the most desired times. The effect can be exploited by manipulating $(\tilde{\alpha}^{(A)}, \tilde{\alpha}^{(B)})$,

proving that Pareto-improving ramp metering schemes exist for single-destination freeways with time-elastic demand. These is further discussed in the final section.

2.3.2 Downstream restrictions (Spillover effect)

Let us now assume that $q_D(t) \leq q_{\max}$, with the possibility of spillovers from link MD. A feasible equilibrium can be found in a sequential manner. Figure 2.8 illustrates the steps for the case of a permanent restriction, $q_D < q_{max}$, but the solution is general for any time-dependent $q_D(t)$.

Step 1: *Aggregate departure curve and maximum delay.* As before, we look for solutions in which the aggregate departure curve D_D is given by the single bottleneck solution with capacity $q_D(t)$ and total population $N^{(A)} + N^{(B)}$; see Figure 2.8a. This defines the duration of the queuing episode, $\Pi = [t_s, t_f]$. By virtue of Property UE(2), necessarily one of the origins flows at both t_s and t_f. We shall assume that A-vehicles discharge uninterruptedly in the interval $\Pi^{(A)} \equiv \Pi$ (i.e., $t_s^{(A)} = t_s$,$t_f^{(A)} = t_f$) and that B-vehicles discharge in an interval $\Pi^{(B)} \subseteq \Pi$ (i.e., $t_s^{(B)} \geq t_s$, $t_f^{(B)} \leq t_f$) possibly with interruptions. Of course, The roles of the origins can be reversed and we show below that this depends on the *population-to-priority* ratio. The equilibrium delay function for origin A, $\tau^{(A)}(t) = \max\{\tau^{(A)}(t), \tau^{(B)}(t)\}$, is obtained from $t_s^{(A)}$ and $t_f^{(A)}$; i.e., $\tau^{(A)}$ coincides with the single origin delays. This can be expressed graphically in conjunction with curve D_D by means of an *A/D curve* $T^{(A)}$ such that the horizontal difference between $T^{(A)}$ and D_D for any departure time t is the A-delay: $\tau^{(A)}(t) = t - {T^{(A)}}^{-1}(D_D(t))$; see Figure 2.8a. Neither D_D nor $T^{(A)}$ are cumulative counts for A-vehicles since D_D gives the cumulative count for A and B-vehicles together.

Step 2: *Delays on the common link.* Since the aggregate departure curve D_D and the maximum delays, $\tau^{(A)}$, are given, we can use the procedure explained in §2.2.2 to obtain the actual delays on link MD, τ_{MD}: first, shift D_D according to (2.4) to obtain the capacity curve at M, D_M^D, and the M-delays, $\nu_{MD}(t)$ (see Figure 2.8b); then, obtain the curve of cumulative flows through M, D_M, as the lower envelope of D_M^D and the higher curve underneath $T^{(A)}$ with slope $\leq q_{\max}$. The horizontal difference between D_M and D_D is the actual delay τ_{MD} (see Figure 2.8c). These delays are highlighted by the shaded area of the figure.

Step 3: *Solution for secondary origin (B).* We first look for a starting time $t_s^{(B)} > t_s = t_s^{(A)}$. For each candidate $t_s^{(B)}$, the equilibrium delays $\tau^{(B)}$ are given. Thus, we can again define a curve $T^{(B)}$ such that the horizontal difference between $T^{(B)}$ and D_D gives directly the equilibrium delay for origin B (see Figure 2.8d). To obtain the outflows from B users consider that: (i) no departures from B can take place when $\tau^{(B)}$ is less than τ_{MD} (i.e., in the non-shaded bands of Figure 2.8d when the curve $T^{(B)}$ dips below D_M), (ii) positive departures rates for B-vehicles always occur when $\tau^{(A)} > \tau^{(B)} \geq \nu_{MD}$ (on the shaded areas in Figure 2.8d), that is, B-vehicles cross the merge when both approaches have queues and, hence, they will flow through the merge using a share $\tilde{\alpha}^{(B)}$ of the capacity. This means that in the shaded intervals $\alpha_D^{(B)} = \tilde{\alpha}^{(B)}$ and outside $\alpha_D^{(B)} = 0$. The total number of B-vehicles passing through D is, therefore, the product of $\tilde{\alpha}^{(B)}$ and the vertical projection of the shaded areas. By changing the starting time $t_s^{(B)}$ we can ensure that the correct number of B-vehicles $N^{(B)}$ is discharged. This is the equilibrium starting time. Figure 2.8e shows the final result. The arrival and departure curves for origin B are just rescaled versions of the A/D curve $T^{(B)}$ and the departure curve D_D: $D_D^{(B)}(t) = \int_{-\infty}^{t} \alpha_D^{(B)}(\epsilon) d_D(\epsilon) d\epsilon$ and $A^{(B)}(t) = \int_{-\infty}^{t} \alpha_D^{(B)}(\epsilon) \dot{T}^{(B)}(\epsilon) d\epsilon$.

Step 4: *Solution for primary origin (A).* Since $\alpha_D^{(A)}(t) = 1 - \alpha_D^{(B)}(t)$, the departure and arrival curve for origin A can be built by scaling curvesD_D and $T^{(A)}$: $D_D^{(A)}(t) = \int_{-\infty}^{t} \alpha_D^{(A)}(\epsilon) d_D(\epsilon) d\epsilon$ and $A^{(A)}(t) = \int_{-\infty}^{t} \alpha_D^{(A)}(\epsilon) \dot{T}^{(A)}(\epsilon) d\epsilon$. Note $\alpha_D^{(A)}(t) = \tilde{\alpha}^{(A)}$ or 1 in agreement with our assumption that $\alpha_D^{(A)}(t) > 0$. This completes the solution.

Three different solution types can arise depending on the populations $\{N^{(A)}, N^{(A)}\}$ and the priority ratios $\{\tilde{\alpha}^{(A)}, \tilde{\alpha}^{(B)}\}$.

Solution 1B. If the number of B-vehicles is small enough so that $t_s^{(B)} > t_s^{(A)}$, the solution is as illustrated by Figure 2.8e. In this case, B-commuters suffer less cost then A-commuters since $|\mathbf{OB'}| < |\mathbf{OA'}|$. This solution arises if $0 < N^{(B)} \leq \tilde{N}^{(B)} = \tilde{\alpha}^{(B)}(N^{(A)} + N^{(B)} - N_U)$, where N_U is the total number of vehicles that find the merge uncongested; see Figure 2.8c.

Solution 1A. A symmetric solution to 1B where A-vehicles experience less cost is obtained by interchanging the superscripts A and B. This solution arises if $0 < N^{(A)} \leq \tilde{N}^{(A)} = \tilde{\alpha}^{(A)}(N^{(A)} + N^{(B)} - N_U)$.

Solution 2. It is also possible to find an equilibrium for the remaining situations with intermediate values of $N^{(B)}/N^{(A)}$. In these cases, $t_s^{(A)} = t_s^{(B)}$ and both origins share the same cost $|\mathbf{OB'}| = |\mathbf{OA'}|$. Consideration shows that the solution is now as in Figure 2.8f. As required by the traffic model, $(\alpha_D^{(A)}, \alpha_D^{(B)}) = (\tilde{\alpha}^{(A)}, \tilde{\alpha}^{(B)})$ in the intervals when both approaches are queued (*shaded area*) and $(\alpha_D^{(A)}, \alpha_D^{(B)})$ is arbitrary in the periods when M is below capacity (*cross-hatched area*). An equilibrium is reached for any $(\alpha_D^{(A)}, \alpha_D^{(B)})$ that generates total discharges matching the populations $N^{(A)}$ and $N^{(B)}$. This solution arises if $\tilde{N}^{(B)} \leq N^{(B)} \leq \tilde{\alpha}^{(B)}(N^{(A)} + N^{(B)})$ or $\tilde{N}^{(A)} \leq N^{(A)} \leq \tilde{\alpha}^{(A)}(N^{(A)} + N^{(B)})$.[6]

[6]Note that a solution always exists since N_U changes continuously with $N^{(A)} + N^{(B)}$ and $k_j \ell$.

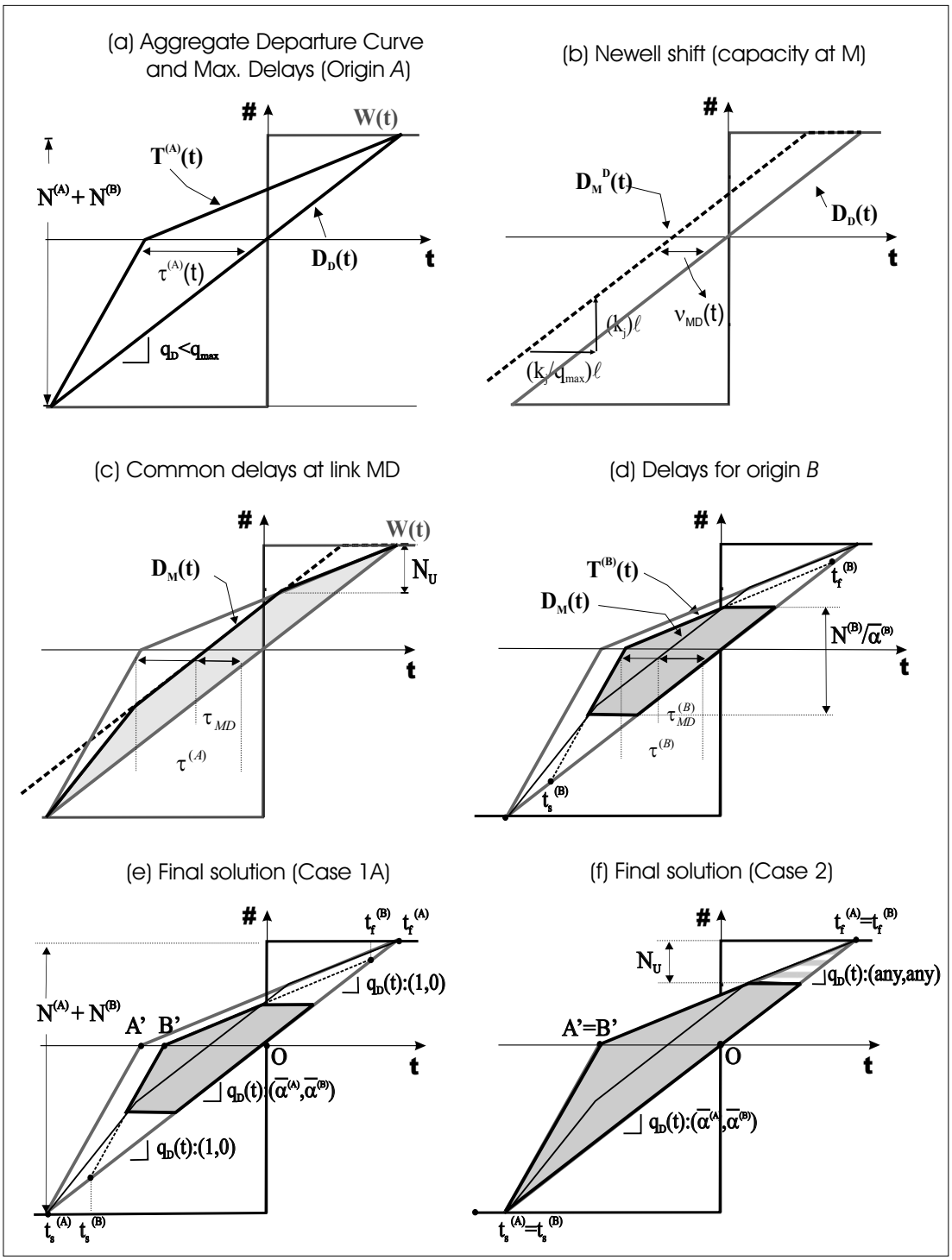

Figure 2.8. Equilibrium solution, queues in link MD (Permanent bottleneck).

The same procedure applies for any time-variable capacity at D. Figure 2.9 shows an example. Note that an equilibrium pattern arises where commuters in B depart during two separated episodes. Figure 2.10 shows the spatial evolution of the queues for time intervals where different queuing patterns arise. Thin black arrows represent the observed flows at the critical sections; thick white narrows, the movements of the head and tail of the queues.Sequence numbers refer to states shown in the cumulative plot.

From an algorithmic point of view, it is important to highlight the following property of the equilibrium solution when the network is homogeneous: *given any time-dependent capacities at D and M, $q_D(t), q_M(t)$, both D_D and D_M can be constructed as function of the total population $N^{(A)} + N^{(B)}$ (and independently of the ratio $N^{(A)}/N^{(B)}$)*, that is, *the equilibrium traffic behavior on link MD is independent of the distribution of populations upstream of M.* This property allows to decompose the problem when more than two origins exists as we shall show in chapter 3.

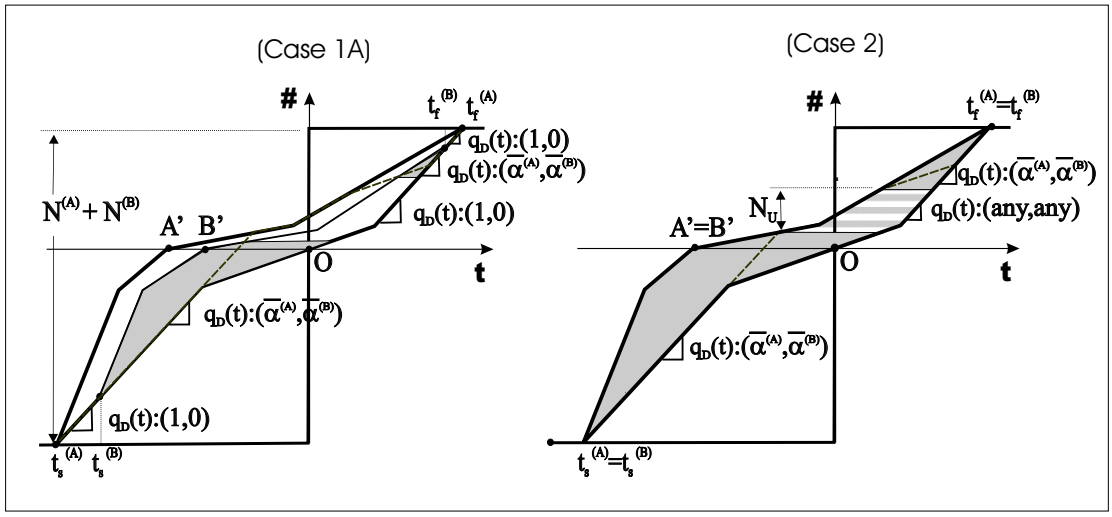

Figure 2.9. Equilibrium solution, queues in link MD (Time-dependent bottleneck).

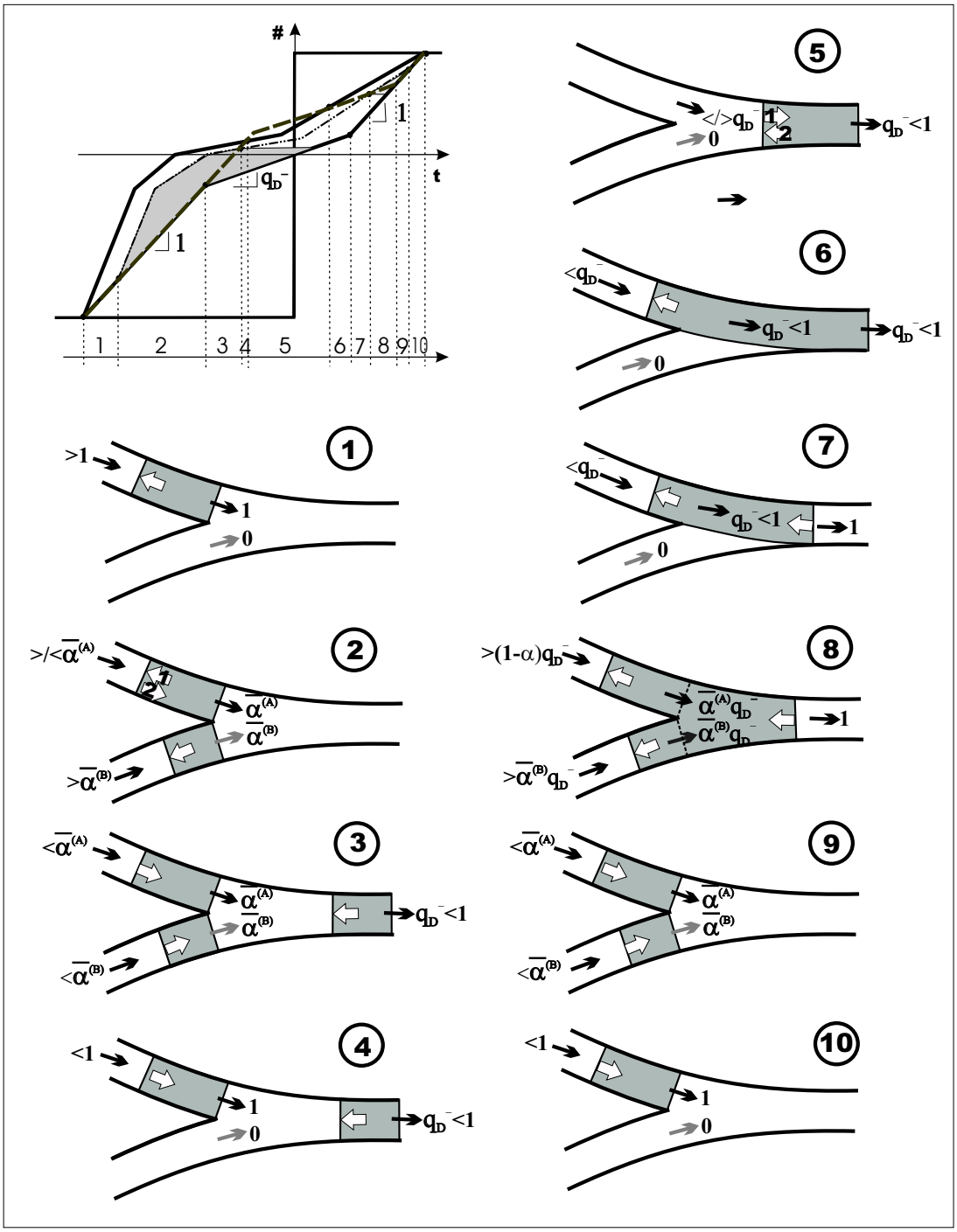

Figure 2.10. Physical queue evolution in equilibrium solution (time-dependent capacity at D).

A comparison of Figure 2.8e and Figure 2.8f with the single origin solution in Figure 2.2 and the solution with merging effects only in Figure 2.7 reveals some interesting insights. In all cases, the population from the origin with the largest *population-to-priority ratio* (A in our case) experiences the same cost equivalent to the cost commuters would suffer in a single origin scenario with total population $N^{(A)} + N^{(B)}$. On the other hand, commuters from the other origin (B) can incur lower cost. Curiously, this cost reduction decreases with the number of vehicles that can be stored in link MD, $k_j\ell$. In the extreme case where $k_j\ell \to 0$, we recover the solution of section 2.3.1 (albeit with a time-dependent capacity) which is actually the least total cost scenario. This is may seem paradoxical at first sight since the provision of extra storage space in link MD makes things worse even though the link capacities remain unchanged (!). The explanation is that the extra space allows A-commuters to mix with B-commuters in the downstream queue. This dilutes part of the segregation advantages that the merge gave to origin B. Again the policy-making implications of these effects are discussed in the final section.

2.4 Departure-time equilibrium and point queue models

In the last decade, a number of works [Akamatsu, 2001; Arnott et al., 1993a; Bernstein et al., 1993; Huang and Lam, 2002; Kuwahara, 1990] have proposed point-queue models with fixed link capacities to solve network equilibrium subject to departure-time choice. These models improve the realism of previous static models by including transient queuing phenomena, but they are still quite restrictive since they neglect the two important forms of link-to-link interaction explicitly consider in our model: *tail-*

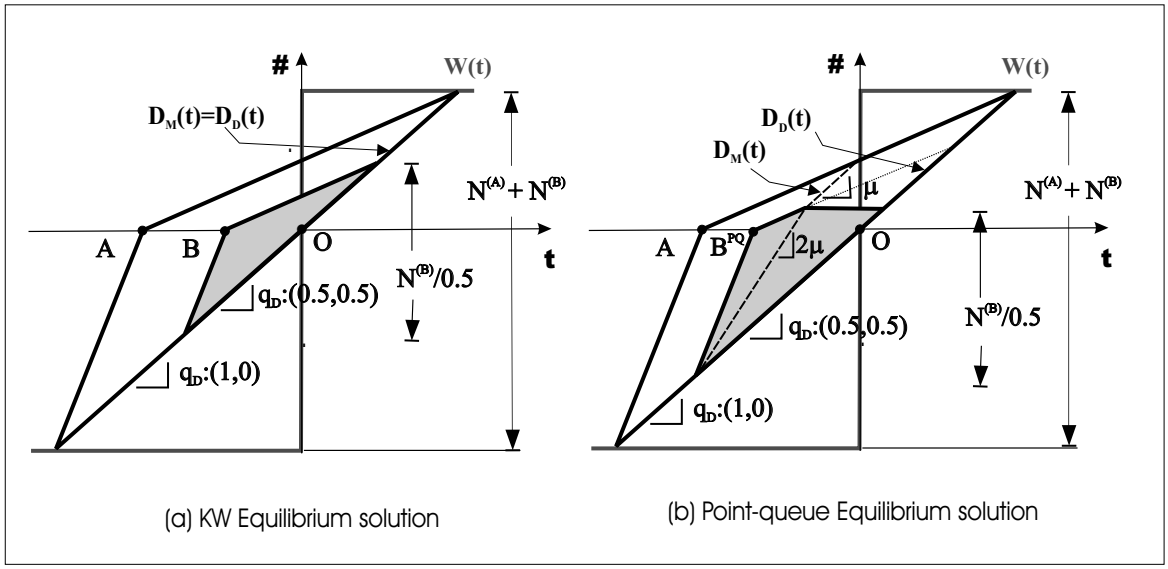

Figure 2.11. KW model vs. point-queue model. No downstream restrictions.

to-head (spillovers) and *head-to-head* (merging competition). These omissions lead to predictions that substantially overstate total cost as shown below. It is therefore important to incorporate link interactions in network models. The goal is achievable because the complexity of the problem appears not to be increased by doing so.

Consider first the issue of over-prediction and start with the constant-capacity case in §2.3.1. Since the above-mentioned point queue models do not restrict merging flows they predict that queues develop at D, although these queues would never appear in reality. The delays produced by these queues must be common to both origins. The ideas of section 2.3.2 can now be used to see that Figure 2.11b is a point-queue equilibrium with queues at D when $d_M(t) > d_D(t)$. This can be compared with Figure 2.11a, which is the solution of section 2.3.1 with $(\tilde{\alpha}^{(A)}, \tilde{\alpha}^{(B)}) = (0.5, 0.5)$. Clearly, the point-queue models predict a significantly larger equilibrium cost for B-users, $|\mathbf{OB^{PQ}}|$, and the wrong location and length of the queues.

The same overstatement of equilibrium delays can be observed if we compare the

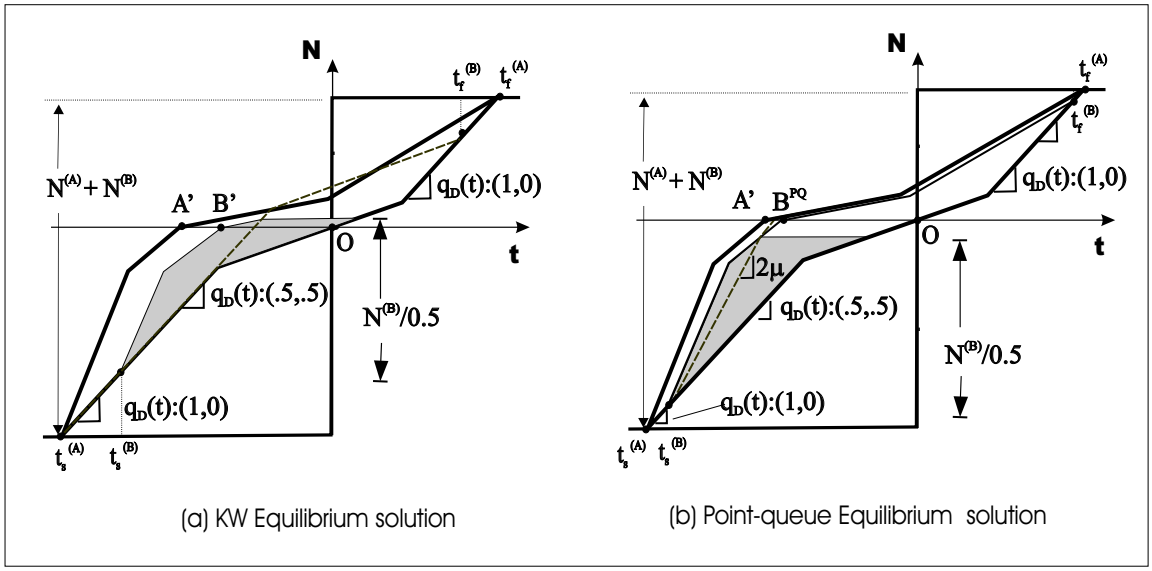

Figure 2.12. KW model vs. point-queue model. Downstream restrictions.

KW solution and the *point-queue* solution when time-dependent restrictions exist at D, as shown in Figure 2.12.

Apart from their realism, solving models with spillover and merge effects involves the same degree of difficulty. The biggest complication with all models (see Figure 2.11) is keeping track of the common delays on link MD to obtain consistent solutions that preserve FIFO. In some cases, it may even be simpler to solve the model with merges and spillovers, since delays turn out to be a function of the downstream conditions and the maximum delays, which can be known a-priori. As we shown in §2.3.2, this knowledge can be used to decouple the equilibrium solution by origin, and to simplify the solution method (this is always the case when the network is homogeneous; for other cases, see chapter 5). The same simplification does not work with fixed-capacity point-queue models, however, because in this case the delays on link MD always depend on the origin-specific flows from upstream.

2.5 Policy implications

It was shown in Arnott et al. [1993a] that total system cost could decrease if one decreases the capacity of a link in a fixed-capacity point-queue network without a route choice. The result is interesting because it suggests that ramp-metering schemes could yield benefits in situations where the conventional wisdom (with fixed departure times) would say that none are to be had. On the other hand, the finding is of limited use because the assumptions underlying rarely arise in ramp-metering practice. Thus, it is fair to ask whether the effect would also arise under the less restrictive assumptions of the model here, and whether it would be any more prevalent. The findings at the end of sections 2.3.1 and 2.3.2 suggest that the answer to both questions should be affirmative since it was shown that if storage capacity is not an issue then a Pareto-improvement can always be achieved by giving some priority to one of the origins. Although this result is only demonstrated for a network where both approach capacities are equal or greater than the capacity at the destination, the result is more general. As shown later in chapter 5 and appendix A for the general case, similar improvements to those of section 2.3.2 can be obtained if the metering rate is constrained never to starve the destination bottleneck for flow. The reason for the generality is that the merge allows the origin flows to interact in a detrimental way, and this happens whether or not the queue-mixing effect identified in Arnott et al. [1993b] also arises. Since in most cases priority should go to the narrower and lower populations approach, the results suggest that contrary to common practice priority in multi-origin freeways should go to the ramps closest to the bottleneck.

It is also shown in §2.3.2 that reducing the storage capacity of link MD (i.e., its length) can reduce delay. This result is just as interesting because it shows that bringing the origins closer to the destination not only decreases free-flow travel time, but

it also decreases delay (!). If the effect continues to arise with multi-origin networks, as we expect, it should have significant policy ramifications because it indicates that the travel costs added by congestion decrease with population density, if one holds the total population constant. This may seem paradoxical because it says that the denser a city the lesser its crowding cost. Next chapter 3 extends the analysis to multi-origin networks and gives more precise answers to these questions.

Chapter 3

A Single Freeway Model

THIS CHAPTER extends the morning commute analysis of chapter 2 to a long freeway leading to a single destination; see Figure 3.1. This model is relevant since a long homogeneous freeway is the logical unit of analysis for mono-centric cities with ring-radial street networks. We characterize the equilibrium solution under departure-time choice and propose an algorithm to solve the problem with the KW model of traffic flow. We show that the freeway problem can be decomposed merge-by-merge and solved recursively as a series of two-origin merge problems (equivalent to the problem in chapter 2). The preliminary insights derived in that chapter about the relationship between population distribution, timing decisions and congestion evolution can now be confirmed through test examples on this network. We show that congestion increases with population sprawl and decreases when greater priority/accessibility is given to downstream origins. These results should have important implications for the treatment of urban congestion in cities with predominantly CBD-bound commute trips.

The chapter is organized as follows. Sections 3.1, 3.2 and 3.3 extend the problem statement of chapter 2 to the freeway setting. Section 3.1 introduces the problem

formulation. Section 3.2 extends the KW traffic model to general single destination network instances. For consistency with the rest of the chapters, the KW model description is kept more general than what it is required to solve the freeway problem. In section 3.3, the departure-time equilibrium conditions and the properties of the solution are revisited. The solution procedure is presented in section 3.4. Numerical solutions are discussed in 3.5 and policy implications discussed in section 3.6.

3.1 The freeway network

3.1.1 Network representation

We shall consider a long freeway running through a residential area (see Figure 3.1). Commuters access the freeway at a number of ramps and travel to a common destination located just downstream of point O. The freeway network can be represented by a graph consisting of a set of N nodes, a set of R origins (a subset of the nodes) and a set of L directed links connecting the nodes. Each origin r is connected by a unique link, (i.e., ramp) to a merging section of the freeway represented by node i (with $i = r$). Merging nodes are located at distances $\{x_i, i = 1 \dots R\}$ from O and are numbered in increasing order of distance; $\ell_{ij} = x_i - x_j$ is the length on link (i, j). A fixed demand $\eta^{(r)}$ travels every from each origin r to the common destination during the morning rush period. For convenience, N_i represents the aggregated population originating upstream of node i including $\eta^{(i)}$ (i.e., $N_O \equiv N_1$ represents that total population).

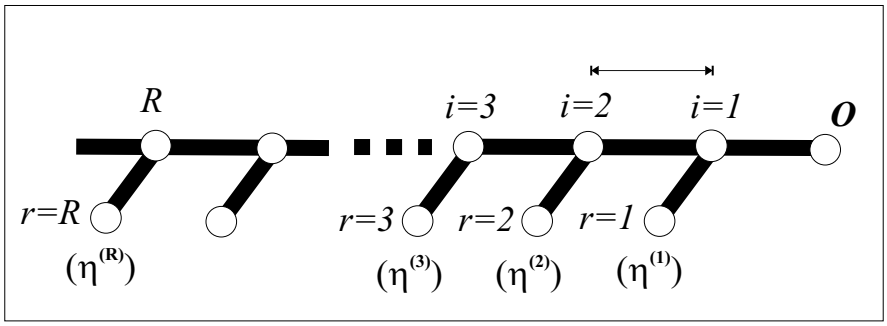

Figure 3.1. Single freeway network.

3.1.2 Traffic representation

Figure 3.2 shows a scheme with the notation used to characterize the traffic solutions (a summary of all the notation used can be found in appendix B). A network assignment is defined by a set of cumulative departure curves from each origin, $\{A^{(r)}(t), i = 1 \dots R\}$ (indexed by the time of desired arrival at the merge node $i = r$). The traffic model solution should then provide as a function of these inflows, the origin-specific cumulative arrival curves at the destination, or outflows at point O, $\{Y^{(r)}(t), i = 1 \dots R\}$ and the origin-specific total travel times experienced $\tau^{(r)}(t)$, that we express as a function of the arrival time to O, t. $A^{(r)}$ and $Y^{(r)}$ are non-decreasing continuous piecewise differentiable functions with respective derivatives (i.e., flows) $a^{(r)}$ and $y^{(r)}$. Under the desirable FIFO property, then

$$A^{(r)}(t - \tau^{(r)}(t)) = Y^{(r)}(t) \tag{3.1}$$

The traffic solution is characterized in each link (i, j) by $A_{ij}(t)$, the cumulative number of vehicles entering the link and $D_{ij}(t)$, the cumulative number of vehicles leaving it (a_{ij} and d_{ij} are the respective time derivatives or flows). To guarantee the assignment feasibility, we must also assume that queues may form upstream of

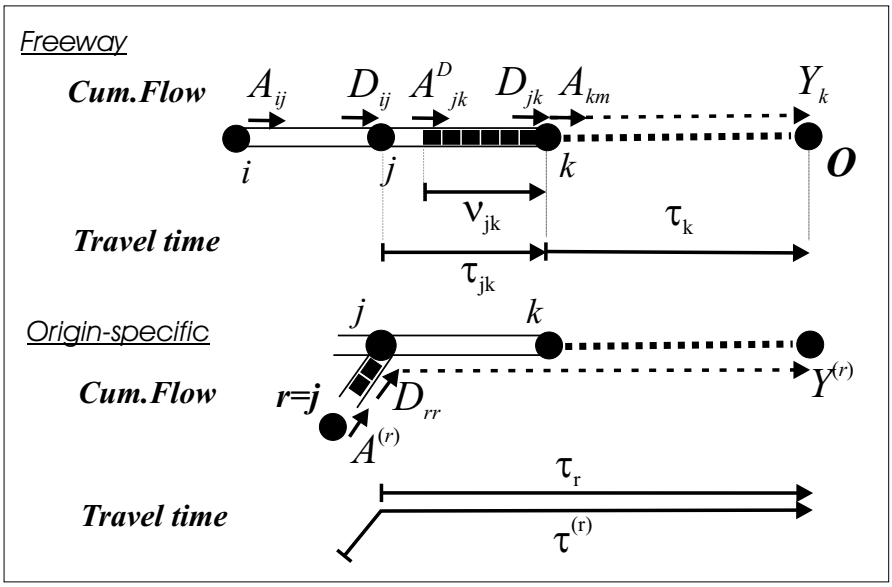

Figure 3.2. Traffic behavior representation and notation.

each network entry point when the available capacity is not sufficient to accept the desired inflow $A^{(r)}(t)$. We do this by modelling ramps as links with infinite storage capacity, whose outflows can be restricted by the inherent ramp capacity or because a freeway queue has backed up beyond the merge. Ramps queues have cumulative inflows $A^{(r)}(t)$ and departure outflows $D_{rr}(t)$, i.e., D_{rr} is the actual cumulative inflow from origin node r into the freeway merge node i ($i = r$), which is controlled by the merge behavior.

Finally, in a single destination network, the traffic solution can be alternatively represented by the following set of functions: $\{Y_i(t), i = 1 \dots N\}$, the cumulative number of commuters crossing node i (i.e., crossing link (i, j)) that have arrived at O by time t, and $\{\tau_i(t), i = 1 \dots N\}$, the travel times from each node i to O experienced by vehicles arriving at time t.[1] Under FIFO conditions, the actual arrival

[1]It is important to distinguish between τ_i, the travel time experienced in the freeway from i to O which is common to all upstream users, and $\tau^{(i)}$ which includes the waiting time in the ramp as well and it is particular to origin-i commuters.

and departure curves at each link are automatically recovered as

$$A_{ij}(t - \tau_i(t)) = D_{ij}(t - \tau_j(t)) = Y_i(t). \tag{3.2}$$

Link travel (i, j) times experienced by vehicles arriving at the destination at time t are $\tau_{ij}(t) = \tau_i(t) - \tau_j(t)$. The set of variables $\{Y_i, \tau_i\}$ is convenient because it represents all traffic information as a function of the arrival time to the destination, t, only.[2] This will become useful to characterize equilibrium solutions in §3.3 and §3.4.

3.2 KW network model

As in chapter 2, we adopt Newell's *simplified kinematic wave (KW)* model of traffic flow to represent traffic behavior [Newell, 1993]. We shall as well neglect the free-flow travel times since these are fixed for each origin and independent of the arrival times; hence, τ_i and $\tau^{(r)}$ represent the delays due to congestion exclusively (the term *delay* or *trip time* will be equivalent through the rest of the chapter). In agreement with this assumption, traffic behavior in each link (i, j) follows a triangular *flow-density diagram* with maximum capacity q_{ij}, jam density k_{ij}, free-flow speed $v_f = \infty$ and a modified backward wave-speed $1/\hat{w} = 1/v_f + 1/w$; see Figure 3.3.

Under the KW model, the entering flows at a link must be determined both from upstream and downstream traffic conditions, since queue spillovers from a downstream node may reduce flow capacity an the entrance of a link. The exact graphical procedure suggested by Newell [1993] and briefly outlined in §2.2.2 can be used to analyze spillovers; see Figure 3.3.

[2]This representation holds for any single destination network with unique O/D paths as long as the traffic model preserves FIFO. In this case, it can be shown that it also holds for many-to-one networks with route choice under user equilibrium conditions [Akamatsu and Kuwahara, 1999].

We shall consider first the case where a unique link (i, j) heads into a link (j, k), with (possibly) different characteristics. In the absence of spillovers, the inflow into link (j, k) would only restricted by the capacity q_{jk}. If A_{ij} represents the cumulative inflow into link(i,j), then the *no-spillover* upstream demand to enter (j, k) would be the highest curve underneath A_{ij} with slope $\leq q_{jk}$ (i.e., the output curve of a point-queue bottleneck with arrivals A_{ij} and capacity q_{jk}). We represent this unrestricted demand as A^U_{jk}. On the other hand, for an actual curve of departures at the downstream end of a link (j, k), D_{jk}, the shifted curve

$$A^D_{jk}(t) \doteq D_{jk}(t - \ell_{jk}/\hat{w}) + k_{jk}\ell_{jk} \tag{3.3}$$

tracks the effects of the backward queuing wave just downstream of location j and represents the maximum cumulative number of vehicles allowed into (j, k) as given by downstream conditions. We shall call A^D_{jk} the spillover curve at link (j, k).

Newell showed that the actual arrival curve at j, A_{jk}, is the lower envelope of A^U_{jk} and A^D_{jk}, i.e.,

$$A_{jk}(t) = D_{ij}(t) = min\{A^U_{jk}(t), A^D_{jk}\}. \tag{3.4}$$

Mathematically, the procedure can also be represented in terms of flows as follows. The maximum possible departure rate from link (i, j), d^{max}_{ij} (i.e., the outflow from link (i, j) in the absence of any downstream restriction) will be equal to the capacity q_{ij} if a queue exists on the link; otherwise, it will be equal to the inflow on link (i, j) as observed downstream of node i; hence,

$$d^{max}_{ij} = \begin{cases} q_{ij} & \text{if } A_{ij}(t) > D_{ij}(t) \text{ (i.e., queue upstream)} \\ a_{ij}(t), & \text{if } A_{ij}(t) = D_{ij}(t) \text{ (i.e., no queue upstream)} \end{cases} . \tag{3.5}$$

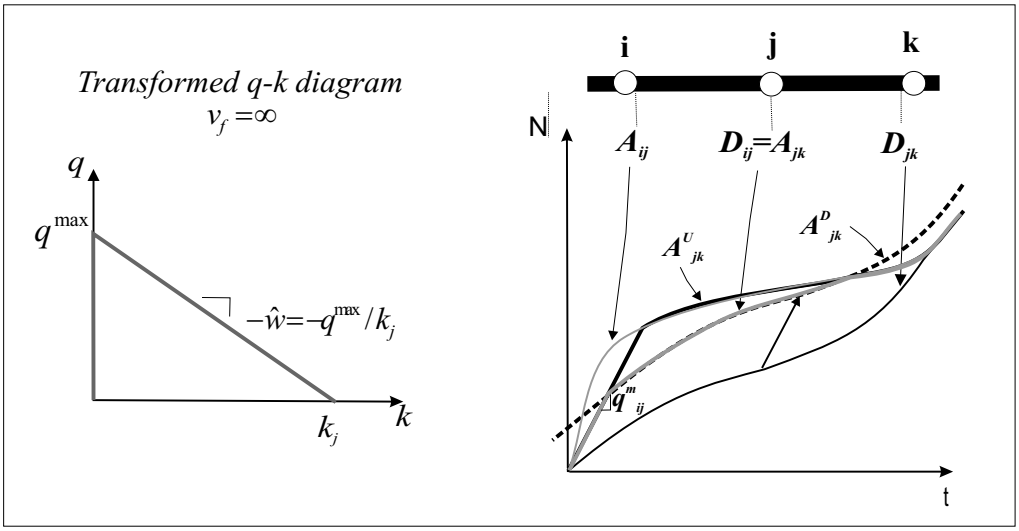

Figure 3.3. KW model: q-k diagram and analysis of spillovers.

Note that $a^U_{jk} = \min\{d^{max}_{ij}, q_{jk}\}$.

The maximum flow allowed to enter link (j,k), $a^{\max}_{jk}$, by the downstream queue is

$$a^{max}_{jk} = \begin{cases} a^D_{ij} \equiv d_{jk}(t - \ell_{jk}/\hat{w}) & \text{if } A_{jk}(t) = A^D_{jk}(t) \text{ (i.e., spillover)} \\ q_{jk} & \text{if } A_{jk}(t) < A^D_{jk}(t) \text{ (i.e., no spillover)} \end{cases} . \tag{3.6}$$

The actual flow at j is given by the minimum of (3.5) and (3.6), i.e.,

$$a_{jk}(t) = d_{ij}(t) = \min\{d^{max}_{ij}(t), a^{max}_{jk}(t)\}. \tag{3.7}$$

In the freeway network, a merge node exists upstream of each link (j,k) with upstream approaches (i,j) and (i',j) (actually one of the approaches is always a origin access ramp). In this case, the outflows in each of the incoming links must share the capacity available downstream according to some priority rules. Following Daganzo [1995a, 1996], we assume that each upstream link (i,j) uses as large a share

α_{ij} of the capacity available downstream as possible subject to the following rules: (i) flow on the approaches can never be larger than the exiting demand d_{ij}^{max} , (ii) the sum of flows can never exceed the (j,k) entering capacity a_{jk}^{max} and (iii) when queues exist on approach (i,j) , its share α_{ij} must equal or greater than a fixed merge-specific share $\tilde{\alpha}_{ij}$, i.e., $\alpha_{ij} \geqslant \tilde{\alpha}_{ij}$ (when queues exist on both approaches, α_{ij} and $\alpha_{i'j}$ must be equal to the merge-specific fixed ratios since it is assumed that $\tilde{\alpha}_{ij} + \tilde{\alpha}_{i'j} = 1$). Mathematically, this maximization can be expressed as a simple linear program:

$$\max\{d_{ij} + d_{i'j}\} \text{ s.t.:} \left\{ \begin{array}{l} 0 \leqslant d_{sj} \leqslant d_{sj}^{max} \; s = i, i'; \; d_{ij} + d_{i'j} \leqslant a_{jk}^{max}; \\ d_{sj} \geqslant \tilde{\alpha}_{sj} a_{jk}^{\max} \text{ if } d_{sj}^{\max} = q_{sj} \; s = i, i' \end{array} \right\} \tag{3.8}$$

where $d_{ij}^{\max}$ (or $d_{i'j}^{\max}$) and a_{jk}^{max} are given by equations (3.5) and (3.6). The solution to (3.8) is

$$d_{sj}(t) = \min\{d_{sj}^{max}(t), \alpha_{sj}(t) a_{jk}^{max}(t)\} \; s = i, i'$$

with

$$\alpha_{ij}(t)\, a_{jk}^{\max}(t) = \left\{ \begin{array}{ll} a_{jk}^{\max}(t) - d_{i'j}^{\max}(t) & \text{if } d_{i'j}^{\max}(t) < (1 - \tilde{\alpha}_{ij}) a_{jk}^{\max}(t) \\ \tilde{\alpha}_{ij} a_{jk}^{\max}(t) & \text{o.w.} \end{array} \right. \tag{3.9}$$

$$\alpha_{i'j}(t)\, a_{jk}^{\max}(t) = (1 - \alpha_{ij}(t)) a_{jk}^{\max}(t)$$

Note that from equation (3.9) the flow observed just downstream of j is $a_{jk}(t) = d_{ij}(t) + d_{i'j}(t)$. Figure 3.4 depicts the possible different states of the merge graphically and the associated flow values.

Finally, at the origin nodes, considering the infinite-storage ramp with arrival curve $A^{(r)}$ and departure curve D_{rr}, the maximum departure demand as given by the queues state in the ramp is

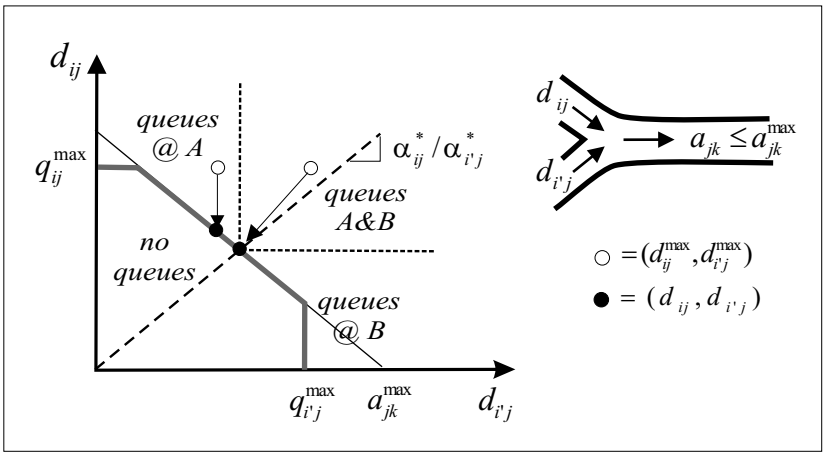

Figure 3.4. General merge behavior. Possible flows depending on merge queuing states.

$$d_{rr}^{max}(t) = \begin{cases} q_{rr} & \text{if } A^{(r)}(t) > D_{rr}(t) \\ \min\{a^{(r)}(t), q_{rr}\} & \text{if } A^{(r)}(t) = D_{rr}(t) \end{cases} \tag{3.10}$$

Hence, the actual departure rate $d_{rr}(t)$ is determined from $d_{rr}^{\max}(t)$ and the share of capacity available for the ramp approach at the merge given by (3.9).

Equations (3.5)-(3.10) allow tracking the evolution of the traffic solution for any time t as a function of any origin inflows $A^{(r)}$. In the departure time equilibrium problem, however, the curves $A^{(r)}$ are unknown and must be determined as a function of the equilibrium conditions.

3.3 Departure-time equilibrium assignment

3.3.1 Departure-time equilibrium definition

A network assignment is a dynamic user equilibrium (DUE) under departure time choice if no commuter can reduce his commute cost by changing unilaterally his departure (equivalently, arrival) time. We adopt the same assumptions as in chapter

2. First, the individual commute cost is defined as a combination of a trip time cost and a scheduled delay arrival penalty. Furthermore, we assume that commuters are homogeneous in their valuations of travel time and schedule delay and have a common preferred arrival time or *deadline* $w = 0$. Expressing cost in units of trip time and referring all the functions to the commuter's arrival time to the destination, the cost for an *origin-r* commuter with deadline w is

$$C^{(r)}(t, w) = \tau^{(r)}(t) + p(t) \tag{3.11}$$

where $\tau^{(r)}(t)$ is the travel delay and $p(\cdot)$ is the common convex *schedule penalty* function, with $\dot{p}(s) < 1$ [Daganzo, 1985; Smith, 1984].

Under the common deadline assumption, the general network equilibrium problem consists of finding a set $\{A^{(r)}, r = 1 \dots R\}$ that will yield through the traffic model outflows $\{Y^{(r)}, r = 1 \dots R\}$ and delays $\{\tau^{(r)}, r = 1 \dots R\}$ (hence, costs $C^{(r)}, r = 1 \dots R$) which satisfy the following set of constraints:

$$y^{(r)}(t)(C^{(r)}(t) - \tilde{C}^{(r)}) = 0 \quad \forall r \forall t \tag{3.12}$$

$$Y^{(r)}(T) - \eta^{(r)} = 0 \quad \forall r \tag{3.13}$$

where $\tilde{C}^{(r)} = \inf_t C^{(r)}(t)$ is the equilibrium cost for origin r, and T is a sufficiently large time denoting the end of the study period. Equation (3.12) states that the cost for commuters on the same origin should be equal for any chosen arrival time and equal or larger for any other non-chosen times. Equation (3.13) represents the population conservation constraint for each origin. Equations (3.12)-(3.13) also express the equilibrium conditions for general networks if one includes route choice in the definitions of $C^{(r)}(t, \text{route})$ and $\tilde{C}^{(r)} = \inf_{t,\text{route}} C^{(r)}(t, \text{route})\ \forall r$.

Using the KW model, no one has yet proven that a set $\{A^{(r)}\}$ verifying (3.12)-(3.13) always exists for general network instances, nor there is an evident strategy to find them in cases where they may exist. In the case of a single freeway, we shall show that an exact analytical procedure can be developed that always yields a solution; therefore, the existence of equilibrium solution is proven by construction. This procedure exploits the same properties UE(1) and UE(2) of the equilibrium used in chapter 2, which are briefly revisited next.

3.3.2 Equilibrium properties revisited

Under a departure time equilibrium where commuters share the same deadline and penalty functions the following properties must hold:

Property UE(1)(*or Parametric representation of equilibrium).* The equilibrium delays (or trip times) for each origin, $\tau^{(r)}$, are uniquely determined by the time of arrival of the first commuter from origin suffering any delay, $t_s^{(r)}$, and so is the equilibrium cost, $\tilde{C}^{(r)}$.

Property UE(2). *(or sequential ordering of delays).* Given an ordering by origin of initial times $t_s^{(r)}$ (or equivalent of cost $\tilde{C}^{(r)}$), the equilibrium trip times for any other departure time follow this same order, i.e., $\forall r, s \tilde{C}^{(r)} > \tilde{C}^{(s)} \Longleftrightarrow \tau^{(r)}(t) > \tau^{(s)}$.

Recall that the equilibrium delays were given by (2.9) as $\tau^{(r)}(t|t_s^{(r)}) = \tilde{C}^{(r)} - p(t) = p(t_s^{(r)}) - p(t)$. This equation holds independently of the arrival flow $y^{(r)}(t)$ being positive or zero. It simply states what the equilibrium delays would have to be if there were some arrivals from origin r at time t.[3]

[3]We can as well define $t_f^{(r)}$ as the time of arrival of the last commuter suffering any delay (which depends on $t_s^{(r)}$ through the relationship $p(t_s^{(r)}) = p(t_f^{(r)})$. Then, $\Pi^{(r)} = [t_s^{(r)}, t_f^{(r)}]$ (with $t_s^{(r)} < w$ and

Property UE(1) suggests, as in chapter 2, that the equilibrium solution be found as a function of the vector of times $\mathbf{t_s} = \{t_s^{(r)}, r = 1 \dots R\}$. This could be done in the following manner: choose a set of $t_s^{(r)}$ which defines uniquely the equilibrium travel times for each origin as a function of the arrival time, $\{\tau | t_s\} = \{\tau^{(r)}(t|t_s^{(r)}), r = 1 \dots R\}$; then, find (by some yet unspecified method) a set of inflows $\{A^{(r)}\}$ that would yield, through the KW traffic model, outflows $\{Y^{(r)}\}$ and delays $\{\tau^{(r)}\}$ equal to $\{\tau^{(r)}|t_s^{(r)}\}$ and such that$\{A^{(r)}, Y^{(r)}, r = 1 \dots R\}$ and satisfy the equilibrium condition (3.12). The total outflows $\{Y^{(r)}(T)\}$ may not match the $\eta^{(r)}$. Hence, we would have to change the set $t_s^{(r)}$ until $\{Y^{(r)}(T)\}$ satisfies (3.13). In practice, this procedure involves two steps: first, solving the *traffic inversion* problem that yields the $\{A^{(r)}\}$ and $\{Y^{(r)}\}$ as a function of $\{\tau^{(r)}|t_s^{(r)}\}$; second, performing a multi-valued search over the $t_s^{(r)}$.[4] These tasks can be streamlined in the case of a linear homogeneous freeway, since, as we show in the next section, both steps of the procedure can be decomposed by origin. Therefore, $t_s^{(r)}$, $A^{(r)}$ and $Y^{(r)}$ can be obtained for each origin sequentially. Before, however, it is necessary to rewrite the necessary condition (3.13) in terms of $t_s^{(r)}, A^{(r)}, Y^{(r)}$.

3.3.3 Necessary conditions

We shall assume that the actual total delays experienced from merge node r (i.e., the access node to the freeway for origin r) $\tau_r(t)$ are known; then equations (3.11) and (3.12) imply

$$y^{(r)}(t)(\tau_r(t) - \tau^{(r)}) \leqslant 0 \quad \forall r \forall t. \tag{3.14}$$

$tsr > w$) gives the time interval where *origin-r* commuters can potentially arrive since outside $\Pi^{(r)}$ the cost is larger, i.e., $y^{(r)}(t) = 0 \; \forall t \in \Pi^{(r)}$.

[4]These tasks are not clearly defined *a-priori*. If they can always be done, the procedure could potentially be applied to any single destination network (with or without route choice).

Furthermore, when the arrival flow $y^{(r)}(t)$ is positive, we can take derivatives in (2.9) and in the FIFO condition (3.1) to obtain respectively the following relationships: $\dot{\tau}^{(r)}(t) = -\dot{p}(t)$, and $a^{(r)}(t - \tau^{(r)}(t)) = y^{(r)}t)/(1 - \dot{\tau}^{(r)}(t))$. Together, these yield the following necessary conditions on the departure and arrival flows

$$\tau_r(t) - \tau^{(r)} \leqslant 0 \Rightarrow a^{(r)}(t - \tau^{(r)}(t)) = y^{(r)}(t)/(1 + \dot{p}(t)) \forall t \in \Pi^{(r)} \tag{3.15}$$

$$\tau_r(t) - \tau^{(r)} > 0 \Rightarrow a^{(r)}(t - \tau_r(t)) = y^{(r)}(t) = 0 \tag{3.16}$$

3.4 Equilibrium solution procedure

We shall assume that the freeway is uniform in width, i.e., all links including the ramps have equal capacity q_{max}, but the ramp priority coefficients $\tilde{\alpha}_{rr}$ can vary across origins.[5] For simplicity, we will use the notation $\tilde{\alpha}_{rr} = \alpha^{(r)}$.

3.4.1 Aggregation-by-merge and recursive logic

We propose a recursive procedure to solve the freeway problem based on the following assumptions:

(1) At the destination node O, the equilibrium departure curve $Y_O(t)$ and maximum equilibrium delays $\tau_O^M(t)$ can be calculated as if we had a single origin problem with total population N_O and a bottleneck capacity q_{max}.

[5]The equal freeway-ramp capacity is formally required for the proposed algorithm to work always. In most solutions, however, it can be shown that equilibrium outflows on each ramp never exceed $\alpha^{(r)}q_{max}$; hence, the solution is only affected by $\alpha^{(r)}$ but not by the capacity of the ramp. One can hence choose a $\alpha^{(r)}$ that reflects the lower capacity of ramps, i.e., if q_r is the actual capacity of the ramps, the $\alpha^{(r)}q_{max} < q_r$.

(2) At each merging node i, commuters can be treated as if they came from two origins (one for each approach) with populations N_{i+1} and $\eta^{(i)}$. The equilibrium problem can be reduced to a two-origin problem, equivalent to that solved in chapter 2 with a known capacity curve at O and a spillover curve at merge node i determined from downstream conditions.

The basic solution procedure works recursively progressing upstream merge-by-merge, using prior solutions $\{Y^{(r)}, r = 1 \dots i-1\}$ to determine both the capacity available at O and the spillover curve at the merge. The details of a step are as follows. First, we define the available capacity curve at O such that $q_i(t) = q_{max} - \sum_{r=1}^{i-1} y^{(r)}(t)$ and the spillover curve at merge i as a function of $\{Y^{(r)}, r = 1 \dots i-1\}$ using (3.3) sequentially at each link $(j, j-1)$ for $j = 1 \dots i$. The problem is then equivalent to the one on chapter 2 (the equivalence between the two-origin and the freeway merge problem is represented in Figure 3.5, the only difference being the way the spillover curve is defined). From the two origin problem solution, we obtain the solution for origin i, $\{Y^{(i)}, \tau^{(i)}\}$.

We shall justify that the procedure actually yields an equilibrium solution, using the following equilibrium property of the homogeneous network.

Property UE(3). (*or aggregation by-node*) For each node i, the equilibrium departure curve from i, $A_{i,i-1}$, and the arrival curve at the destination of commuters passing i, Y_i, depend on the total upstream population and the distribution of downstream populations, but not on the distribution of commuters across upstream origins.

Proof: Property UE(3) follows directly from the insight derived in chapter 2. We showed that in the two origin homogeneous network problem, for any given time-dependent capacities at the destination node D, $q_D(t)$, and at the merge M, $q_M(t)$,

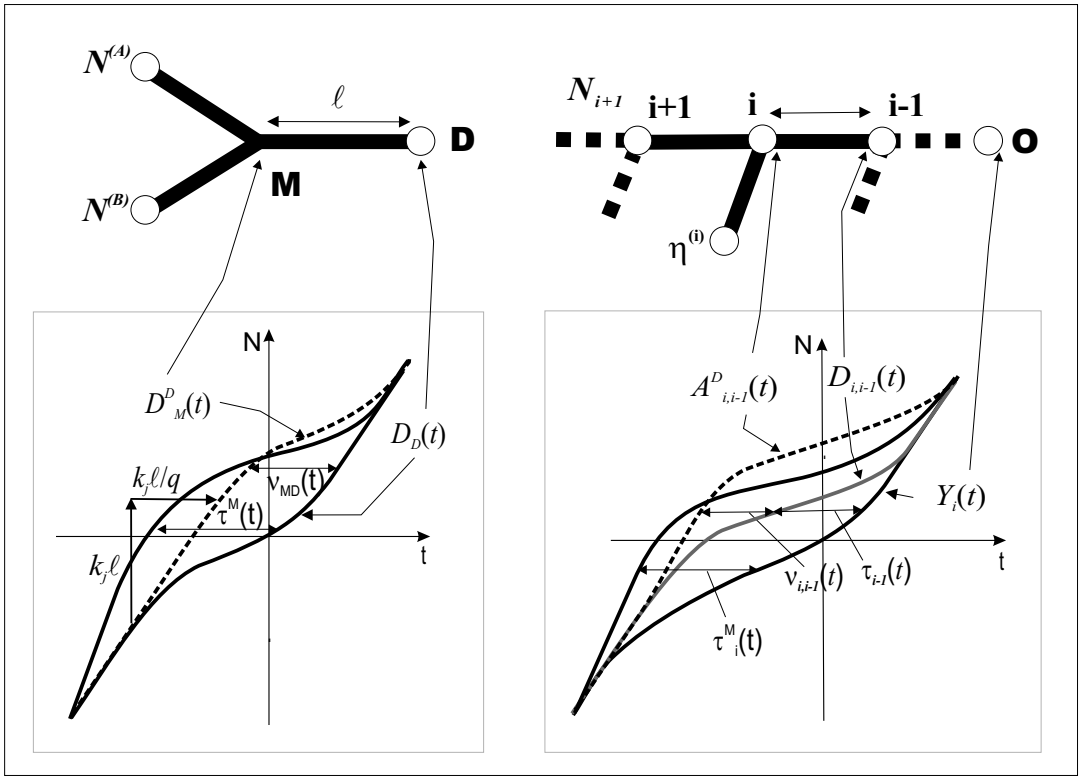

Figure 3.5. Equivalence: 2-origin problem vs. Merge-i problem.

an equilibrium solution existed where the actual departure curves from D, $D_D(t)$, and from M, $D_M(t)$, were a function of the total population $N^{(A)} + N^{(B)}$ only but did not depend on the ratio $N^{(A)}/N^{(A)}$. (Recall that this was mainly a consequence of Property UE(2)). The same principle applies now. Consider first the equilibrium solution for the last two origins R and $R-1$ whose routes coincide at merge $R-1$. For any possible solution at the downstream origins $\{Y^{(r)}, r = 1 \ldots R-2\}$, the equilibrium problem for origins R and $R-1$ reduces to a two-origin network problem with known available capacity at O , $q_{R-1}(t)$, and at merge $R-1$, $a^{\max}_{R-1,R-2}$; hence, the equilibrium arrival curve at the destination for people crossing merge $R-1$, Y_{R-1} and the observed departure curve at node $R-1$, $D_{R-1,R-2}$ are only a function of the aggregated population upstream of node $R-1$, $N_{R-1} = \eta^{(R)} + \eta^{(R-1)}$. That

is, property UE(3) holds for merge node $R-1$. It is apparent too that commuters at downstream origins will then 'act' as if a single origin existed upstream of merge $R-1$. Hence, the equilibrium solution for these origins will be as if origin R and $R-1$ where combined, i.e., as if the freeway had $R-1$ origins. Hence the foregoing logic can now be applied to the shortened freeway to show that the property UE(3) holds for $r = R-2$ and then by recursion to the rest of origins. ■

Note in particular, that, $Y_O(t) \equiv Y_1(t)$ will be a function of the total population N_1 and the freeway capacity q_{max} only. Knowledge of $Y_1(t)$ allows determining, by solving a three link problem: $\{Y^{(1)}(t), \tau^{(1)}(t)\}$, which is the solution for origin 1 and $\{Y_2(t), \tau_1(t)\}$, which gives the upstream population departures curves from O and from merge node $i = 1$. In turn, the latter curves allow finding $\{Y^{(2)}(t), \tau^{(2)}(t)\}$ and $\{Y_3(t), \tau_2(t)\}$ by solving another 3-link problem, etc... The specifics of this procedure are described below.

3.4.2 Solution algorithm

The algorithm is summarized in table 3.1. The recursion to obtain $\{Y_{i+1}(t), \tau_i(t)\}$ from $\{Y_i(t), \tau_{i-1}(t)\}$ is implemented more efficiently by also including the state variable τ_i^M, the virtual delay for the aggregated upstream of node i (which includes $\eta^{(i)}$), and iterating instead on $\{Y_i, \tau_{i-1}, \tau_i^M\}$. It turns out that τ_i^M represents the worst case delays or delays experienced from the most disfavored origins upstream of i (we call these maximum delays). Recall from chapter 2 that necessarily one of the two upstream origins experienced these worse delays (equal to the delays experienced if all upstream population came from a single origin) while the delay for the other origin could be less. Now too, the delays for some upstream origins can be less than τ_i^M.

At each iteration, equilibrium behavior is link $(i, i-1)$ is directly obtained from

$\{Y_i, \tau_{i-1}, \tau_i^M\}$ (*step 1*). The departure curve upstream of node $i-1$ is directly given by $D_{i,i-1}(t-\tau_{i-1}(t)) = Y_i(t)$. To obtain the arrival curve at link $(i, i-1)$, $A_{i,i-1}$ (i.e., the actual departure curve at node i), we can use directly the procedure in §3.2. Since by virtue of Property UE(3), we treat all upstream users as coming from a single upstream origin, the curve $A^M_{i,i-1}(t-\tau_i^M(t)) = Y_i(t)$ represents the desired inflows into link $(i, i-1)$; hence, the demand curve $A^U_{i,i-1}$ is defined as the highest curve with slope $\leq q_{max}$ underneath $A^M_{i,i-1}$. The spillover curve $A^D_{i,i-1}$ is defined as a function of $D_{i,i-1}$ by (3.3). Then, the actual $A_{i,i-1}$ is $A_{i,i-1}(t) = \min\{A^U_{i,i-1}(t), A^D_{i,i-1}(t)\}$.

The merge i equilibrium solution (*step 2*) is obtained with the recipe of chapter 2. We look for the solution on the origin $\bullet$ with lower *population-to-priority* ratio, for which potentially $\tau^{(\bullet)} < \tau_i^M$ (i.e., we choose origin i if $\eta^{(i)}/\alpha^{(i)} < N_{i+1}/(1-\alpha^{(i)})$; the virtual origin M, otherwise). Only a minor variation to the solution is introduced. Recall that when $\eta^{(i)} > \tilde{N}^{(i)}$ – where $\tilde{N}^{(i)} = \alpha^{(i)}(N_{i+1} + \eta^{(i)} - N_i^U)$ represents the maximum population from origin i that can depart when the merge is congested – then $\tau^{(r)} = \tau_i^M$ and the excess commuters $\eta^{(i)} - \tilde{N}^{(i)}$ from origin i cross the merge at any arbitrary rate at times when the merge is not at capacity (i.e., they arrive at O when $\tau^{(r)}(t) = \tau^M_{i+1}(t) = \tau_i(t)$). To avoid choosing an arbitrary flow rate when this happens, one can assume procedure the excess $\eta^{(i)} - \tilde{N}^{(i)}$ commuters from origin i will depart from the virtual origin M instead. This does not alter the equilibrium solution since this portion of commuters always flows when neither of the approaches is congested. Obviously a symmetric modification applies when the virtual origin M is the approach with the lower population-to-priority ratio.

Table 3.1. Departure-time equilibrium algorithm for the homogeneous freeway

Let Φ denote the operation that solves the single bottleneck problem for a population N and a capacity $q(t)$, possibly time-dependent, i.e., $\{Y, \tau\} = \Phi(N, q(t))$ where $Y(t)$= equilibrium departure curve and $\tau(t)$ = equilibrium delay.

1. *Initialization* ($i = O$). Solve a single bottleneck with population N_1 and capacity q_{max} to obtain $Y_1(t) = D_{10}(t)$ and $\tau_1^M(t)$, i.e., $\{Y_1, \tau_1^M\} := \Phi(N_1, q^{\max}\}$. Set $\tau_O(t) := 0$ and update node: $i = 1$

2. Solve link $(i, i-1)$ equilibrium behavior. The substeps are:

   2.1 Departure curve, $D_{i,i-1}$, given by $D_{i,i-1}(t - \tau_{i-1}(t)) = Y_i(t)$.

   2.2 Arrival curve, $A_{i,i-1}$, given by $A_{i,i-1}(t) = \min\{A^U_{i,i-1}(t), A^D_{i,i-1}(t)\}$, where: $A^D_{i,i-1}$ from (3.3) and, $A^U_{i,i-1}(t)$ is given by the higher curve with slope $\leqslant q^{\max}$ underneath $A^M_{i,i-1}(t - \tau_i^M(t)) = Y_i(t)$.

   2.3 Obtain from $\tau_i(t) = t - A^{-1}_{i,i-1}(Y_i(t))$

3. Solve the merge problem with populations $\{N_{i+1}, \eta^{(i)}\}$, aggr. departure curve $Y_i(t)$ and delay from node i, $\tau_i(t)$. Obtain $\{Y^{(i)}, \tau^{(i)}\}$ and $\{Y_{i+1}, \tau^M_{i+1}\}$. The substeps are:

   3.1 Choose the approach $\blacksquare$ with minimum *population-to-priority* ratio (i.e., $\bullet$= origin i if $N_{i+1}/(1 - \alpha^{(i)}) > \eta^{(i)}/\alpha^{(i)}$

   3.2 Solve equilibrium for the approach $\bullet$ according to recipe of chapter 2:

   $$\{Y_\bullet, \tau^M_\bullet\} := \Phi(\min\{N_\bullet, \tilde{N}^{(\bullet)}\}, y_i^{(\bullet)}\}, \text{ where}$$
   $$y_i^{(\bullet)}(t) = \begin{cases} \tilde{\alpha}_\bullet y_i(t) & \text{if } \tau^{(\bullet)}(t) > \tau_i(t) \\ 0 & \text{if } \tau^{(\bullet)}(t) \leqslant \tau_i(t) \end{cases}$$

   3.3 Update solution in each approach:

   $$Y^{(i)} = Y_\bullet,\ \tau^{(i)} = \tau^M_\bullet$$
   $$Y_{i+1} = Y_i - Y_\bullet,\ \tau^M_{i+1} = \tau^M_i$$
   (The opposite is true if approach $\bullet$ is the upstream freeway)

4. Update node: $i = i + 1$. Repeat 1-3 until $i = R$.

3.5 Numerical analysis

Results are presented for the case where a population N_T is uniformly distributed along an homogeneous freeway of capacity q_{max}, jam density k_j and length ℓ_T. There are R equally-spaced ramps with equal priority ratio $\tilde{\alpha}^{(r)} = \alpha$; hence, $\eta^{(r)} = N_T/R \; \forall r$. The 6 parameters that control this *symmetric* setting $(q_{max}, k_j, L, N_T, \alpha, R)$ can be reduced to 3 if we use as units of time $1/q_{max}$, units of length L and units of population N_T (i.e., $\hat{q} = q/q_{max}$, $\hat{\ell} = \ell/\ell_T$ and $\hat{N} = N/N_T$). The new parameters are: $\hat{k}_j = k_j \ell_T / N_T$, the relative storage capacity of the freeway as a percentage of total population; R, the number of ramps; and α, the ramp priority. Note that $\hat{k}_j$ is inversely proportional to the average population density, N_T/ℓ_T, hence an indirect measure of population sprawl.

The commute cost for each origin, $\tilde{C}^{(r)}$, is expressed as a fraction of the equivalent equilibrium cost suffered in a single origin scenario.[6] Note that the cost $\tilde{C}^{(r)}$ is purely associated with congestion and scheduled delay since it does not include the time portion corresponding to free-flow trip times.

Figure 3.6 shows typical commute cost curves as a function of the origin distance to the destination for different $\tilde{k}_j$ (and fixed $\alpha = 0.2$, $R = 15$). For all three curves, cost increases as a function of distance, the increase being sharper the further away from the destination; beyond a critical location, cost remains constant. The reasoning behind this behavior has to be found in the effects of queues and merge priorities. The principal queues will appear at the most downstream merge and will spill beyond the access ramps located close to the destination during most of the peak time. Commuters in these origins hold an advantage since they can arrive at the preferred

[6] Assuming a V-shaped schedule penalty function with with earliness and lateness conversion rate e and L, the single bottleneck cost is $\frac{N_T}{q_{max}} eL/(e+L)$.

times and save a great part of the delay by cutting ahead of the queue (i.e., they experience some delay at the ramp but this is much less that the equivalent delay experienced in the upstream freeway for the same arrival time). As origins get further away, spillovers are less prominent and commuters choose to arrive at least preferred times where the delays experienced at the ramp are still smaller (instead of withstand the maximum equilibrium queues in the freeway at the most preferred arrival times). Beyond a critical location, commuters in the upstream origins must endure the full delay (either at the ramp or at the freeway) independently of the time they choose to arrive at. This is the same delay they will endure if everybody came from a single origin. Examples of origins in each of these scenarios are shown in Figure 3.7. For each origin, the total equilibrium delay (solid line) and the freeway delay (dashed line) are compared as a function of the arrival time to the destination. The difference, when positive, is the actual delay experienced at the ramp. (On the other hand, when negative, commuters at the origin do not depart.) Note that for the first origin shown $r = 3$ ($\hat{\ell} = 0.08$), freeway delays are never large enough as to discourage arrival in the times close to the deadline. For origin $r = 10$ ($\hat{\ell} = 0.38$), however, commuters are 'forced' to arrive at very early arrival times when the freeway delays are shorter than the equilibrium ones; at the times close to the deadline, delays in the freeway are larger and no flow of commuter from this origin happens. Finally, for origin $r = 15$ ($\hat{\ell} = 0.58$), commuters are indifferent among any arrival time, since they must endure the maximum equilibrium delays either part at the ramp and the freeway or entirely at the freeway.

Consequently with the above behavior, commuting cost should grow as the relative storage capacity increases since spillovers will be less likely. This is confirmed in Figure 3.6. Since storage capacity increases with population dispersion, – recall that

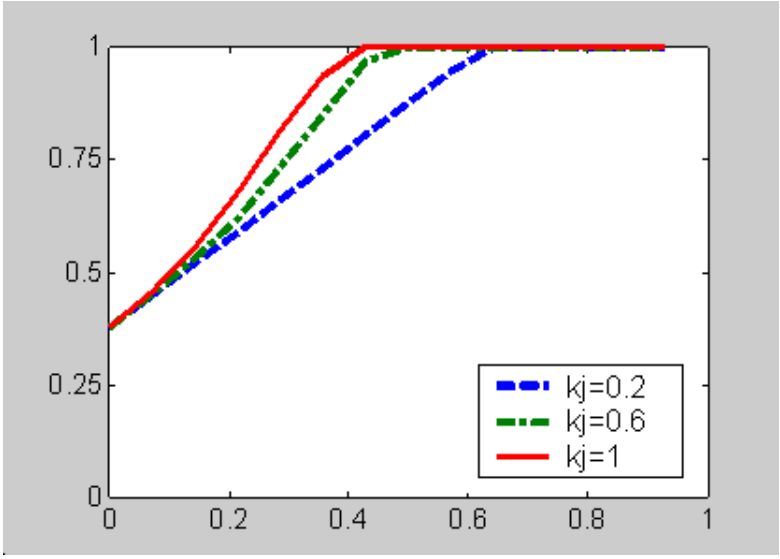

Figure 3.6. Individual congestion cost vs. origin location. Sensitivity with $\hat{k}_j$ ($\alpha = 0.2$; $R = 15$).

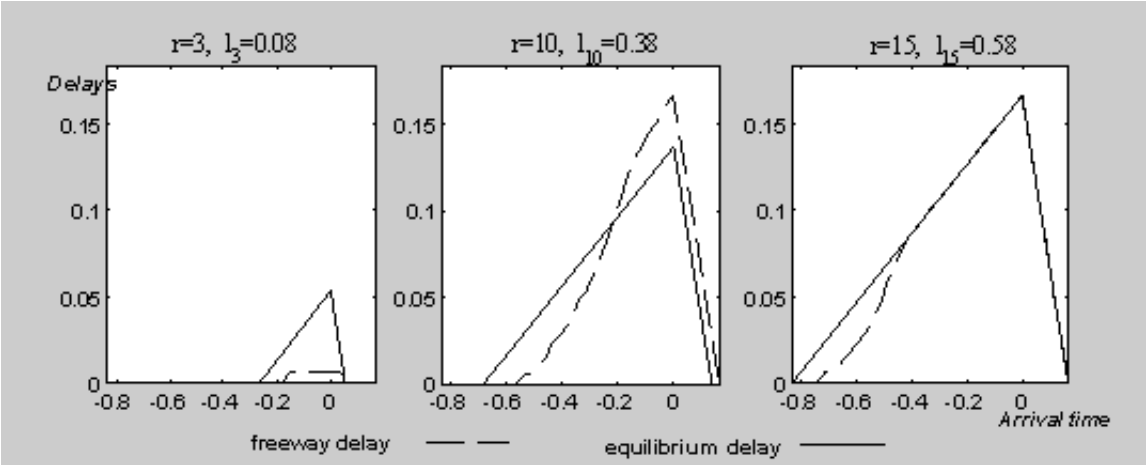

Figure 3.7. Equilibrium delay vs. freeway delay ($\hat{k}_j = 0.6$; $\alpha = 0.2$; $R = 15$).

$\hat{k}_j = k_j \ell_T / N_T$ –, this is implies that a more disperse population results not only on longer non-congested trip times but also on increased congestion in the freeway. This confirms the initial insights derived for the two-origin network in 2.

Conversely, as shown in Figure 3.8, an increase on the number of ramps available renders the individual cost down. The gain is more important in origins closer to

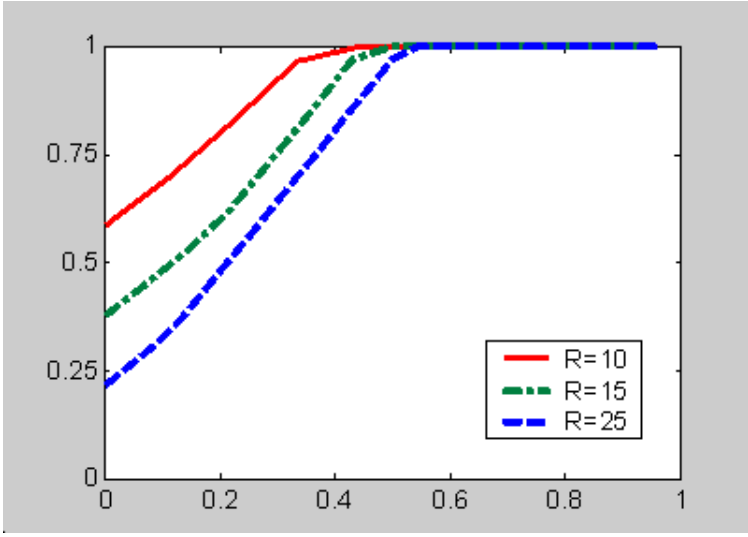

Figure 3.8. Individual congestion cost vs. origin location. Sensitivity with R ($\hat{k}_j = 0.6$; $\alpha = 0.2$).

the destination since these origins benefit from larger outflows during the intervals where commuters can cut ahead of the queue. Figure 3.9 shows that an equivalent effect is obtained if the priority of ramps is increased. This confirms the idea that current ramp metering practice (which conversely reduces priority of downstream access points at the most congested times) may lead to undesirable outcomes.

To assess quantitatively the importance of these phenomena, table 3.2 displays the change on total system cost as a function of both population sprawl and ramp accessibility (cost are expressed as a percentage of the total equilibrium cost in a single origin scenario). The percentage changes indicate that the effects of population sprawl are very limited. Most costs saving arise mainly as a result of the priority that commuters in downstream approaches hold. This somehow hopeful for society since it implies that adequate ramp metering (or equivalently, other access-control policy such as tolls) may compensate for the undesirable sprawl effects.

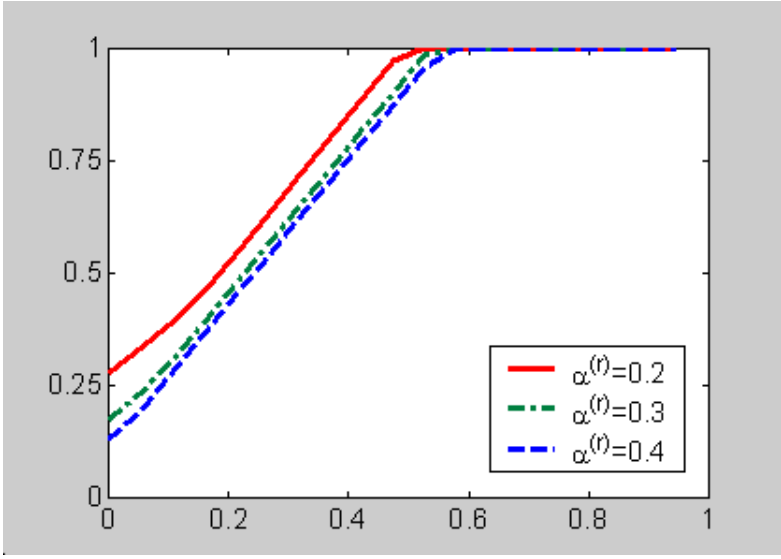

Figure 3.9. Individual congestion cost vs. origin location. Sensitivity with α ($\hat{k}_j = 0.6$; $R = 15$).

Table 3.2. Total cost sensitivity (as % of single origin cost)

	Relative storage capacity ($\hat{k}_j$)				
Ramps	0.2	0.4	0.6	0.8	1
r = 10	0.81	0.82	0.84	0.85	0.86
r = 15	0.73	0.76	0.79	0.81	0.81
r = 20	0.69	0.73	0.76	0.78	0.79
r = 25	0.67	0.71	0.75	0.77	0.77

3.6 Final remarks

The analysis in this chapter should have an important influence on the way urban economists treat congestion. The traditional economic analysis, as discussed in the introductory chapter, has been based on steady-state models that are flawed when applied to spatial/dynamic settings. Here we have presented an alternative robust

characterization of traffic that can be used to study spatial problems more adequately. Since timing decisions and traffic dynamics are determined endogenously, our cost curves provide a consistent structural relationship between population distribution and congestion cost.

Our results suggest that, after all, population distribution may not be such an important driver of congestion costs on corridors with predominantly single destination work trips. In such instances, policies aimed at changing commuters timing behavior (e.g., congestion tolls or access control policies) may be more efficient in general than those aimed at controlling urban growth. Today, there is a renewed interest among policy makers in access-control schemes that combine time-dependent traffic tolls with restricted automobile access. The efficiency gains that can be achieved by across-the-board tolls, however, have a regressive effect on non-wealthy population groups-and this often makes them politically unpopular. More equitable tolling schemes that exempt from the toll a different subset of the vehicular population each day (e.g., on odd or even days, by weekday, etc.) have been devised [Daganzo and Garcia, 2000]. Our equilibrium results strongly suggest that spatial equity should also be considered in the design of tolling schemes, since the desired effects of tolls may differ with location given the distributed impact of queues. In that regard, the design of time and location-dependent tolling strategies to reduce congestion is a promising alternative.

The study of urban location pattern as a function of congestion should be also adapted to our representation of costs. Next chapter embeds the freeway model on a more general urban traffic model. Some simplifications on the traffic assumptions are also considered that yield continuous closed-form solutions and, at the same time, preserve most of the realism of the models.

Chapter 4

Mono-centric Cities

This chapter extends the equilibrium analysis to consider the evolution of congestion in mono-centric cities; i.e., in cities with all trips bound to a central location where all the economic activity happens (the central business district or *CBD*). Commuters can choose to travel to the CBD using the freeway network or an alternative dense grid of city streets. We model this choice under departure time equilibrium. Section 4.1 presents a general model. Customarily, we assume the freeway as subject to congestion and the alternative street network as non-congested (note that we could alternatively consider another non-congested mode of transportation). This general model is too complex and requires numerical solutions. In section 4.2, a radially symmetric city is considered where the analysis can be reduced to study traffic behavior in a single linear freeway/arterial system. We show that the same logic in chapter 3 can be used to solve this linear problem analytically and obtain additional insight. Finally, section 4.3 proposes a continuum approximation. By considering a continuous solution, we are able to obtain an approximate closed-form expression that links congestion cost with the continuous population distribution.

4.1 General case: Freeway network/street grid

A typical mono-centric city transportation network is represented in Figure 4.1. It consists of two main components: a freeway system and an underlying dense grid of streets that allows commuters to travel between each origin and destination, and between freeway access ramps. The freeway system is modelled as a many-to-one tree network with homogeneous links (i.e., with unique paths from each origin to the common destination). The street network is modelled as a continuum with no congestion, with fixed travel times between each pair of points always larger than the equivalent free-flow travel times on the freeway between the same points.

Origins are located along the freeway network. Each origin i represents a neighborhood with population $\eta^{(i)}$. Commuters always prefer to use the freeway from their respective origins when the latter is not congested. As congestion builds up, however, they have the following possibilities: (a) travel from their origin to O on the freeway, (b) use the street network to access the freeway at other less congested origins and travel on the freeway from there to O, and (c) travel directly to O on the street network.

4.1.1 Homogeneous freeway tree network

If we consider that commuters travel to CBD using exclusively the freeway network from their respective origins, one needs only considering the departure time equilibrium problem on the freeway network as a function of the populations $\{\eta^{(i)}, i = 1 \dots R\}$. It should be obvious from Property UE(3) in chapter 3 that the sequential procedure in §3.4 can also be applied to any tree network of homogeneous freeways. In this case, at each merge node, virtual upstream origins must be considered on both approaches to each merge. The solution still progresses traversing each branch of the

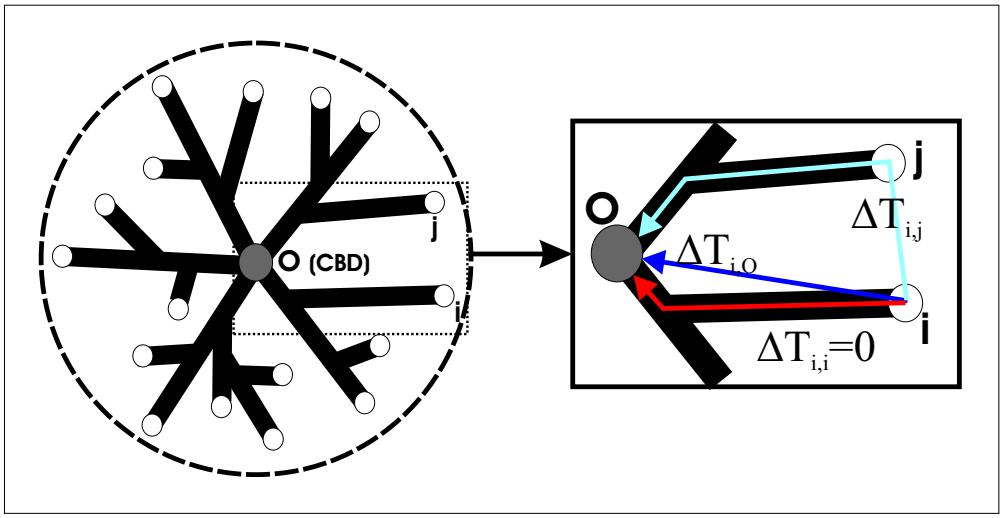

Figure 4.1. Mono-centric city: freeway-street grid representation.

tree, starting from the root.

4.1.2 Freeway network and street grid (Route-choice)

The equilibrium problem with freeway/street route choice can be treated as if we had an elastic demand across origins. Besides commuter timing, we also need to decide for each origin population $\eta^{(i)}$: (a) which portion uses the freeway from origin i, $\eta_i^{(i)}$, (b) which travels through any of the other origins $\eta_j^{(i)}, j \neq i$, and (c) which travels on the street network to the destination, $\eta_0^{(i)}$. Obviously, $\eta^{(i)} = \sum_{j=0}^{R} \eta_j^{(i)}$.

For each origin i, let $\Delta T_{ij} \geqslant 0, j \neq i$ represent the additional time required to travel from origin i to O through origin j assuming free-flow on the freeway (i.e., option (b)) and ΔT_{i0} the additional time required to travel from i to O on the street network (i.e., option (c)).[1] Logically, $\Delta T_{i0} \leqslant \Delta T_{ij} + \Delta T_{j0}$. As in chapter 3, we set the free-flow travel times on the freeway equal to zero. Hence, the ΔT_{ij} can

[1]Note that $\Delta T_{ij} = T_{ij}^S + T_{j0}^F - T_{i0}^F \neq \Delta T_{ji}$ where T_{ij}^S is the travel time between i and j on the streets and T_{ij}^F on the freeway. Equivalently, $\Delta T_{i0} = T_{i0}^S - T_{i0}^F$.

be interpreted as extra travel times (i.e., delays) for the transformed $delay - based$ problem. Furthermore, under the no street congestion assumption, commuters using the street network always arrive on time; hence, their total congestion cost is fixed and equal to ΔT_{i0}. In this case, the population split vector $\boldsymbol{\eta} = \{\eta_j^{(i)} : i = 1 \dots R,\ j = 0 \dots R\}$ must be determined to satisfy the following cost constrains:

$$[\Delta T_{i0} - \tilde{C}^{(i)}]\, \eta_0^{(i)} \leqslant 0 \quad \forall i \tag{4.1}$$

$$[\Delta T_{ij} - (\tilde{C}^{(i)} - \tilde{C}^{(j)})]\eta_j^{(i)} \leqslant 0 \quad i, j = 1, \dots R\,, j \neq i \tag{4.2}$$

Equation (4.1) states that for each origin i, the equilibrium congestion cost in the freeway cannot exceed the additional travel time in the street network since, in this case, commuters would shift to the streets in a proportion that makes freeway congestion cost equal to street excess travel time. Similarly, equation (4.2) states that the difference in congestion cost between two origins must never exceed the additional inter-origin street travel time, since, if this happens, a portion of commuters from the higher cost origin i will choose to travel through the lower cost origin j.

The general freeway/street grid problem can then be treated as a bi-level equilibrium problem:

- At the *lower level*, the freeway network equilibrium problem can be solved conditional on the population split vector $\boldsymbol{\eta}$. Note that the number of commuters travelling in the freeway from each origin will be $\tilde{\eta}_i = \sum_{r=1}^{R} \eta_i^{(r)}$. We can use the algorithm in §3.4.2 to obtain the unique freeway congestion cost associated with each origin.

- At the *upper level*, one must find the vector $\boldsymbol{\eta}$ that satisfies the set of inequality constrains (4.1)-(4.2), using the implicit relationship between $\boldsymbol{\eta}$ and $\{\tilde{C}^{(i)} : i =$

$1, \ldots R\}$ given by the freeway equilibrium problem (i.e, the *lower level*).

This bi-level equilibrium problem can be easily recast as a finite-dimensional nonlinear complementary problem (NCP) or a fixed-point problem (FPP) and solved numerically with one of the several heuristic methods available for this type of problems; see, for instance, Facchinei and Soares [1995] for NCP; Nagurney and Zhang [1998] for projected FFP. Not much insight can be derived in this way and we shall not pursue the procedure forward. Instead, we concentrate on the case of a radially symmetric city.

4.2 Symmetric case: Linear freeway/arterial

In many urban analyses, it is customary to assume a symmetric ring/radial network of freeways and streets and a symmetric population distribution around the city center (or CBD). The problem can then be reduced to study just the behavior in one of the radial freeways. In this case, we can represent the freeway/street network by a system composed of a single linear freeway and a parallel arterial. We shall show that one can make use of the recursive logic in §3.4 to determine the final equilibrium.

4.2.1 Extended solution procedure

Let $\Delta T_{i(i-1)}$ represent the additional required (free-flow) travel time on the street network to move from ramp i to $i-1$ and ΔT_{i0} the additional travel time in the street network to reach the destination O from origin i. We assume that $\Delta T_{i0} = \Delta T_{(i-1)0} + \Delta T_{i(i-1)}$; i.e., travel in the street network is parallel to the freeway. Under this assumption the following property must hold:

Property UE(4). If upstream origin i sends flows through origin k (i.e., $\eta_k^{(i)} > 0$)

then it can also send flow through all the intermediate origins (i.e., $\eta_j^{(i)} > 0,\ j = k \dots i$).

Proof: The proof is immediate by contradiction. Consider that property UE(4) does not hold for an intermediate origin j, that is, origin i sends flow through k but cannot through j. Hence, condition (4.2) implies $\tilde{C}^{(i)} < \Delta T_{ij} + \tilde{C}^{(j)}$ and $\tilde{C}^{(i)} = \Delta T_{ik} + \tilde{C}^{(k)}$, or equivalently combining the two conditions, $\Delta T_{ik} - \Delta T_{ij} < \tilde{C}^{(j)} - \tilde{C}^{(k)}$. Since the parallel travel in the arterial requires $\Delta T_{jk} = \Delta T_{ik} - \Delta T_{ij}$, we get $\Delta T_{jk} < \tilde{C}^{(j)} - \tilde{C}^{(k)}$ which clearly violates condition (4.2). ■

The same recursive logic in §3.4 can be used to determine the $\{\eta_j^{(i)}, i = 1 \dots R, j = 0, \dots, i\}$ and the final equilibrium cost $\{\tilde{C}^{(i)}, i = 1 \dots R\}$. We start by the most downstream merge and solve each merge progressing upstream (i.e., adding one origin at a time). At each merge i, we assume that $\{Y^{(r)}, \tau^{(r)}, r = 1, \dots, i-1\}$ are known and that the associated $\{\tilde{C}^{(r)}, r = 1, \dots, i-1\}$ satisfy (4.1) and (4.2). (Note that now $\{Y^{(r)}, \tau^{(r)}\}$ represent the arrivals and equilibrium delays of all commuters entering the freeway at ramp r; these may include commuters from other origins upstream of r. At the same time, $\tilde{C}^{(r)}$ represents both the equilibrium cost for origin r and the commuting cost experienced by travelling in the freeway from ramp r). For every merge, the solution is obtained in two major steps:

First, we obtain the solution for the *two-origin* problem with virtual upstream origins but without considering route choice for these origins (i.e., $\Delta T_{i,j} = \infty, \Delta T_{M,j} = \infty,\ j = 0, \dots, i$). This can be done as in §3.4.2; steps 1 and 2 in table 3.1. We obtain the solution for origin i $\{Y_{NR}^{(i)}, \tau_{NR}^{(i)}\}$ and the cost $\tilde{C}_{NR}^{(i)}$, where the subscript NR indicates that this solution does not incorporate route choice.

Second, we must update the origin i solution to take in account route choice. By

virtue of property UE(4), we just need to identify the lower j s.t. $\Delta T_{ij} + \tilde{C}^{(j)} < \tilde{C}^{(i)}_{NR}$. Three different scenarios may arise:

(a) $\tilde{C}^{(i)}_{NR} < \Delta T_{i(i-1)} + \tilde{C}^{(i-1)}$. Then $\tilde{C}^{(i)}_{NR}$ satisfies both (4.1) and (4.2) for all $j = 0, \ldots, i-1$. This implies that all commuters from i choose the freeway at ramp i and $\{Y^{(i)}_{NR}, \tau^{(i)}_{NR}\}$ is the origin-i equilibrium solution (at the end of the merge-step i).

(b) $\Delta T_{i(k-1)} + \tilde{C}^{(k-1)} < \tilde{C}^{(i)}_{NR} < \Delta T_{ik} + \tilde{C}^{(k)}$. Then, $\eta^{(i)}_j > 0 \ \forall j = i, \ldots, k$. In this case, some commuters from i choose to use ramps i to k. As shown in chapter 2, a shift in the relative populations between two origins will only increase cost for users in the least cost origin. Hence, with a shift of population from i to $(i-1)$ the equilibrium cost at i must remain $\tilde{C}^{(i)} = \tilde{C}^{(i)}_{NR}$ while the new commuting cost for using the freeway from origin $(i-1)$ must be $\tilde{C}^{(i-1)} \rightarrow \tilde{C}^{(i-1)} = \tilde{C}^{(i)}_{NR} - \Delta T_{i(i-1)}$. Since $\tilde{C}^{(i-1)}$ has changed, condition (4.2) does not hold for $(i-1)$ anymore. Using the same reasoning, $\tilde{C}^{(i-2)} \rightarrow \tilde{C}^{(i-2)} = \tilde{C}^{(i-1)} - \Delta T_{i-1(i-2)}$, and so on, for the rest of downstream origins until origin k.

(c) $\Delta T_{i0} < \tilde{C}^{(i)}_{NR}$. Then, $\eta^{(i)}_0 > 0$, that is, some commuters from i travel to O on the street network. The equilibrium cost for all downstream origins must be updated as $\tilde{C}^{(j)} \rightarrow \tilde{C}^{(j)} = \Delta T_{j0}, \ j = 1 \ldots i$.

The $\{Y^{(r)}, \tau^{(r)}\}$ must then be recalculated for all the origins $r = k, \ldots, i$ for which $\tilde{C}^{(r)}$ has been updated. This can be done using steps 1 and 2 in table 3.1, noting that now the equilibrium $\{t^{(r)}_s, \ r = k \ldots i\}$ are already given by $t^{(r)}_s = p^{-1}(\tilde{C}^{(r)})$ and the equilibrium delays by $\tau^{(r)}(t) = \tilde{C}^{(r)} - p(t)$. This finalizes the merge i solution. By

construction the new $\{Y^{(r)}, r = 1 \dots i\}$ satisfy the equilibrium conditions (4.1) and (4.2). One can then proceed then by iteration to solve the origin/merge $(i + 1)$.

4.2.2 Some basic results

Some important qualitative insights can already be gained even without explicitly solving the problem.

A main conclusion is that the existence of an alternative to the freeway would not always relieve congestion, i.e., it will not always reduce system cost. For instance, in case (b) above when some users have an incentive to use downstream ramps to travel to the CBD but not to travel directly to the destination through the street network, the system congestion cost always increases with respect to the *freeway-only* solution. Note that this situation always arises in cases where the final access to the CBD can only be done through the freeway (e.g, a bridges is the only access to a peninsula or an island like in San Francisco, Manhattan or Hong Kong).[2] This apparent paradoxical behavior is a dynamic reminiscence of Braess's Paradox [Braess, 1968]. Similar paradoxes in a dynamic scenario with fixed time-dependent OD demand and point queues can be found in Akamatsu and Kuwahara [1999]. Our results here are stronger since timing decisions are also considered. From a policy standpoint, this suggests that such undesirable *route-choice* must be discouraged.

In cases when a part of the upstream population uses downstream ramps but also reaches the destination directly through the street network case (case (c) above), there is a cost trade-off: cost savings arise because a net amount of population is completely shifted away from the freeway, but some additional cost is also imposed into downstream users by part of the upstream commuters still using downstream

[2] In this case, $\Delta T_{i,O} = \infty$, $\Delta T_{i,j} < \infty$.

ramps. It is reasonable to believe that normally the former will offset the latter, but there is no evidence that this would always be the case.

In the extreme case when all upstream commuters travel to the destination either using the freeway from their origin ramp or directly through the street network – e.g., the alternative to the freeway is another mode of transportation instead of the street network – [3], then cost always decreases with respect to the *freeway-only* scenario.

4.3 A continuum approximation analysis

To study the behavior of a large system like a city traffic network, it is convenient to consider a continuous model, as an approximation to the (discrete) network formulation. This approximation characterizes the system by means of a few significant continuously distributed variables (e.g., population density) and enables to generalize the analysis to more complicated problems.

The urban economics literature uses extensively the continuous representation to analyze the interactions between congestion, infrastructure provision and residential location in mono-centric cities; see chapter 1, §1.1.4. Unfortunately, these urban models represent traffic behavior and congestion incorrectly by assuming a local congestion model. This literature would benefit from an alternative continuous approximation, based on the freeway equilibrium model proposed in chapter 3.

4.3.1 Traditional vs. KW-based equilibrium representation

We shall first briefly revisit the traditional continuous model as presented in Mills and de Ferranti [1971] and Solow [1973] to compare it with a continuous version of

[3] In this case, $\Delta T_{i,O} < \infty$, $\Delta T_{i,j} = \infty$

our freeway equilibrium solution. In agreement with this literature, a single freeway model (symmetric city) is considered and no route choice is explicitly modelled.

4.3.1.1 Traditional *local congestion* model

A total population N_T is continuously distributed over a length ℓ_T with a density of $\eta(x)$ commuters per kilometer, where x is the distance from the CBD.[4] Every commuter accesses the freeway at his residential location x. Freeway characteristics are given by the freeway width at each location which determines the capacity $q_{max}(x)$. For consistency with the rest of the analysis, a homogeneous capacity $q_{max}(x) = q_{max}$ is considered from now on.

The objective is to obtain an expression of the commuting cost $\tilde{C}(x)$ which in this case is only a function of the travel time from location x, $\tau(x)$. Congestion is modelled assuming a steady-state situation, where travel time per unit length at location x, $\dot{\tau}(x) = \frac{\mathrm{d}\tau(x)}{\mathrm{d}x}$ (i.e., the inverse of the speed), is a function Ψ of the ratio between the location steady traffic flow $d(x)$ and the freeway capacity. The function $\Psi(\centerdot)$ can take several forms but usually, a BPR-like power function is adopted; i.e.,

$$\Psi(x) = \tau_u \left(1 + \frac{d(x)}{q_{max}}\right)^a \tag{4.3}$$

where τ_u represents the free-flow travel time per unit length and a is a dimensionless constant.

Some implicit assumptions about commuter departure timing are adopted to express the traffic volume at each location as a function of the population distribution $\eta(x)$. Customarily, it is assumed that commuters depart uniformly in time so as to joint upstream commuters as they pass through the origin in question. Hence, every

[4]The CBD is assumed not to take up space.

arrival interval contains the same mixture of origins (i.e., all origins) in the arrival stream [Solow, 1973].[5] In this case,

$$d(x) = \nu \int_x^{\ell_T} \eta(x)dx. \tag{4.4}$$

where ν is dimensional factor to account for the departure spread of the population over the rush hour (i.e, ν converts populations into steady flows and has units of $time^{-1}$; it is normally chosen to bound the maximum travel cost to a desired value).

Since the local congestion model does not explicitly consider propagation of congestion in space or time, the assumptions about how people lump together govern completely the shape of the location-based delay function. Unfortunately, these assumptions are arbitrarily adopted more for the convenience of modelling that for their realism.[6] Consequently, any analysis of location patterns based on this local representation of congestion must be considered very carefully. Consider, for example, the case of a uniformly distributed population $\eta(x) = \eta$, a congestion function Ψ with a=1 and neglect free-flow travel times; hence, (4.4) yields $\tau(x) = \kappa\eta(\ell_T x - x^2/2)$ with $\kappa = \tau_u\nu/q_{max}$ and it can be concluded that spreading the same population over double the distance (i.e., $\tilde{\eta} = \eta/2; \tilde{x} = 2x$) will double the congestion cost. This result is, however, only a fictitious consequence of the assumption that all upstream population lump together in the traffic stream. Different assumptions will yield different results.[7]

[5]Users depart in the same interval at all origins except for a time shift equal to the freeway trip time.

[6]Mills/Sollow bundling of all upstream population in a common cohort may be adequate to represent congestion in public transportation systems. One can imagine that a common vehicle picks up people as it travels to the CBD and riders experience a progressive discomfort as the vehicle becomes crowded. However, this assumption is not amenable to model freeway congestion.

[7]For instance, Yinger [1993] assumes that users stagger their departures so as to arrive to the destination in order of distance to the destination. The model assumes a two-dimensional city and accounts for the perpendicular travel time/distance necessary to reach the single freeway. For the linear case above, however, the model yields $d(x) = \nu\eta(x)$ and, assuming $a = 1$ and $\eta(x) = \eta$,

4.3.1.2 Continuous KW-based equilibrium model

A more robust continuous representation of the congestion cost as a function of distance and population distribution can be obtained taking the limit of the discrete-ramp freeway equilibrium solution in chapter 3 as the number of ramps $R \to \infty$. For consistency, one must keep the aggregated ramp priority per unit length constant to a value $\alpha(x)$.[8] Although no formal proof is provided here, extensive simulations showed that the cost curve $\tilde{C}^{(r)}$ converges to a continuous solution $\tilde{C}(x)$ as $R \to \infty$. $\tilde{C}(x)$ is always an upper-bound to the actual discrete-ramp costs with the same ramp priority per unit length. The bound is tight for moderately large number of ramps ($R > 15$).

Using a similar dimensional analysis as in §3.5, a full range of curves representing congestion cost (expressed in units of N_T/q_{max}) as a function of distance (expressed in units of ℓ_T) can be obtained for any population distribution in terms of two main parameters: $\tilde{k}_j = k_j N_T/\ell_T$ and $\alpha = \alpha(x)$. One can as well consider a parameterized family of population distributions, e.g., $\eta(x) = \eta e^{-\beta x}$, and define the cost curves as a function of the new parameter (β).[9]

From a practical point of view, however, it even more convenient to have a simple analytical expression of the cost. An analytical expression can be incorporated in other models of urban behavior (e.g., the mentioned models of residential location choice in Solow [1973]). Unfortunately, an exact closed-form solution is not readily

$\tau(x) = \kappa\eta x$. Hence, the location-based congestion cost turns out to be independent of city size. Yinger's assumption is rather absurd: it basically implies that users do not get in the way of each other; that is, there is no congestion interaction among populations at different distances, which is precisely what one wants to model. It is mentioned here, however, to make evident the dependence of the solution on the arbitrary timing assumptions.

[8] One should consider $\alpha(x)$ such that $\alpha(x) = \lim_{R\to\infty}\left\{\lim_{\Delta x\to 0}\left(\sum_{i\in[x,x+\Delta x]}\alpha^{(r)}/\Delta x\right)\right\}$. The function $\alpha(x)$ has units of $length^{-1}$.

[9] If we normalize to $N_T = 1$ and $L = 1$, then $\eta = \frac{\beta}{1-e^{-\beta}}$.

available for the general case. For that reason, we shall consider some special cases for which exact closed-form solutions can be obtained and derive an approximate expression for the rest.

4.3.2 Full ramp priority solution

It is assumed here that ramp flows can block completely the freeway; i.e., $\tilde{\alpha}^{(r)} = 1$ in the discrete case, $\alpha(x) = \infty$ in the continuous case. Obviously, this is not fully realistic but captures in the limit the fact that downstream ramp users enjoy a priority over upstream traffic and hold a location advantage. If the number of ramps per unit length is high, the solution with this assumption could coarsely approximate real behavior.

The analysis of how cost changes with location is notably simpler with this assumption. Note that commuters at any origin can always flow out of their ramp at the maximum capacity q_{max} whenever the freeway is not blocked immediately downstream. Therefore, in an equilibrium, they choose the most favorable time, no queues form on the on-ramps and all delay occurs in the freeway. Furthermore, the equilibrium cost $\tilde{C}(x)$ must necessarily be non-decreasing with x.[10] The cost $\tilde{C}(x)$ defines interval $[t_s(x),t_f(x)]$ where commuters at x can depart, where $\tilde{C}(x) = -et_s(x) = Lt_f(x)$. Since arriving at $t_s(x)$ implies always cutting ahead of upstream commuters and facing no traffic downstream, one only needs deciding if the infinitesimal population at $[x, x + dx]$, $dN(x) = \eta(x)dx$ departs on the neighborhood of $t_s(x)$ or at other times by comparing the earliness with the delay experienced in any point in $[t_s(x), t_f(x)]$. These delays will depend on the behavior of queues.

[10]Note that if the equilibrium cost for a upstream commuter arriving at time t is smaller than the cost experienced by commuters at some origin downstream, the latter would reduce their cost by cutting ahead of the former.

To provide a clearer understanding, we shall describe the solution for three different scenarios: (a) zero freeway storage ($k_j = 0$), (b) infinite freeway storage($k_j = \infty$), and (c) finite freeway storage ($0 < k_j < \infty$). Rather than providing a formal proof we shall give a graphical intuition for the shape of the cost curve in each of these cases; see Figure 4.2. (Of course, one can always check that the solution proposed is actually exact by using a continuous version of the algorithm in §3.4.2.)

Zero freeway storage. This is equivalent to assuming that sudden capacity restrictions at any section propagate to all upstream sections of the freeway immediately; i.e. vehicle sizes or queues are infinite. In this case, vehicles cannot queue on the freeway and commuters at any ramp have always the incentive to arrive at the closest arrival times to the deadline not used by commuters in downstream ramps; i.e., they behave as if they faced a bottleneck with zero capacity during the interval used by downstream users $[t_s(x), t_f(x)]$ and the full capacity q_{max} otherwise (see Figure 4.2a). Therefore,

$$d\tilde{C}(x) = \frac{dN(x)}{q_{max}} \frac{eL}{e+L}. \tag{4.5}$$

Hence, the equilibrium cost can be directly obtained as

$$\tilde{C}(x) = \frac{N(x)}{q_{max}} \frac{eL}{e+L}. \tag{4.6}$$

Infinite freeway storage. The opposite limit case assumes that queues only have a local effect at every merging section and do not affect upstream ramps capacity. In this case, late departure always implies mixing in common queues with the upstream population. Two different regime may exist:

(a) Commuters originating downstream of a given location x^* (yet to be defined)

always prefer to depart on the early side since otherwise they would lose their location advantage. This implies that

$$d\tilde{C}(x) = \frac{dN(x)}{q_{max}}e. \tag{4.7}$$

(b) Commuters originating upstream of location x^* arrive all mixed during the late interval and hence experience the same cost equal to the maximum single-origin bottleneck cost, $\frac{N_T}{q_{max}}\frac{eL}{e+L}$.

To see how this comes about in more detail consider the equilibrium representation in Figure 4.2b. Users downstream of x^* choose to arrive during $[t_s(x^*), 0] = [\frac{N(x^*)}{q_{max}}, 0]$ ordered by origin distance. Upstream commuters behave as if they they came from a single origin and faced a bottleneck of capacity 0 during $[N(x^*), 0]$ and q_{max} otherwise; see the upper triangle in Figure 4.2b. Note then when commuters at any $x < x^*$ arrive in the early $t_s(x)$, they cut ahead of the upstream demand and experience no queues downstream; hence, their cost is $\tilde{C}(x) = e\frac{N(x)}{q_{max}}$. If they choose to depart late at $t_f(x)$, they will have to mix with upstream commuters which have departed earlier and suffer a cost $\tilde{C}(x) = \frac{N_T}{q_{max}}\frac{eL}{e+L}$. Hence, as long as $eN(x) \leq N_T\frac{eL}{e+L}$, they prefer to depart early and the location x^* is such that $N(x^*)e = N_T\frac{eL}{e+L}$. The final cost curve can be defined by parts in the following manner:

$$\tilde{C}(x) = \begin{cases} e\int_0^x \frac{\eta(x)}{q_{max}}dx = \frac{N(x)}{q_{max}}e & \text{if } 0 \leqslant x \leqslant x^* \\ \frac{N_T}{q_{max}}\frac{eL}{e+L} & \text{if } x^* \leqslant x \leqslant L \end{cases}. \tag{4.8}$$

Finite freeway storage. The finite storage capacity can be intuitively derived as a mix of the two previous cases. Note that in the infinite case, users at $x < x^*$ are discouraged from departing late because all upstream commuters can get ahead of

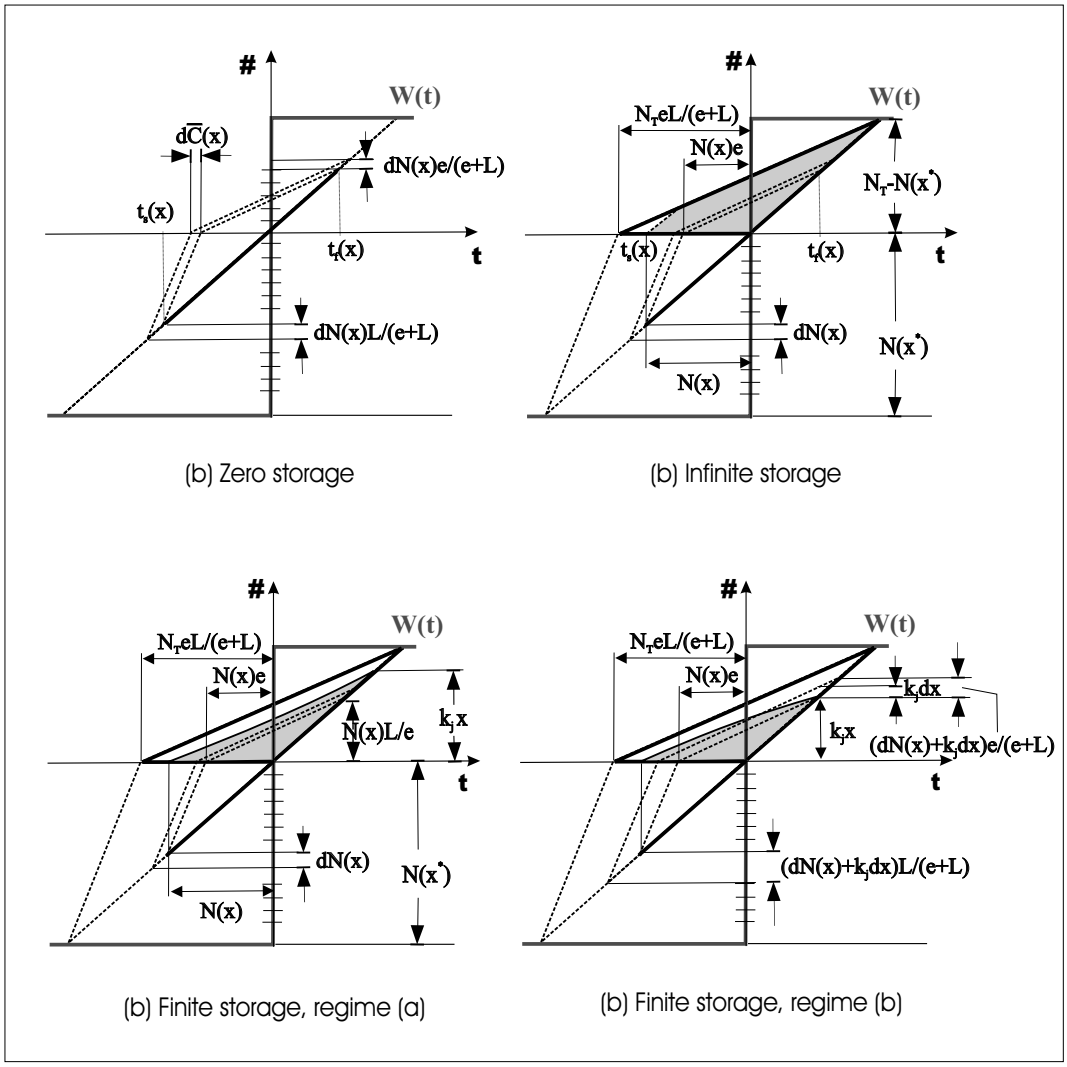

Figure 4.2. Continuous solution. Full ramp priority ($q_{max} = 1$).

them and, hence, they would endure the full delay (or the full lateness once the queue has cleared). With finite storage, there is a maximum quantity $k_j x$ of upstream commuters which can queue in front; hence, if the freeway stops being blocked at $t = 0$ commuters at x can wait to depart at $t = \frac{k_j x}{q_{max}}$ when all the vehicles in front have cleared and the freeway opens. In this case, they experience no delay and suffer a maximum lateness $L\frac{k_j x}{q_{max}}$; see Figure 4.2c-d, where the shaded area represents the

maximum delays experienced at location x given that all downstream commuters departed in the early side. Hence, people can still choose to arrive late and enjoy a location advantage in some cases. Three regimes are possible:

(a) Commuters at x still arrive at the early time $t_s(x)$ only when $N(x) < \frac{L}{e}k_j x$; see Figure 4.2c. Hence, as in the infinite storage case,

$$d\tilde{C}(x) = \frac{dN(x)}{q_{max}}e. \tag{4.9}$$

(b) Commuters depart early and late as soon as $N(x) = \frac{L}{e}k_j x$; see Figure 4.2d. In this case, further observation reveals that if $dN_s(x)$ and $dN_f(x)$ are the shares of the population $dN(x)$ that arrive at $t_s(x)$ and $t_f(x)$ respectively, then $e\frac{dN_s(x)}{q_{max}} = L\frac{(dN_f(x)+k_j dx)}{q_{max}}$;[11] hence,

$$d\tilde{C}(x) = \frac{dN(x) + k_j dx}{q_{max}}\frac{eL}{e+L}. \tag{4.10}$$

Note that for $dN_f(x)$ to be positive, we need $dN(x) > \frac{L}{e}k_j dx$; otherwise, regime (a) will happen.

(c) Finally, as soon as $\tilde{C}(x) = \frac{N_T}{q_{max}}\frac{eL}{e+L}$, users suffer always the maximum cost and are indifferent among any late available arrival time.

For the sake of simplicity, we have discussed a particular case where regime (a) first happens downstream and then (b) follows upstream. This may not always be the case and, depending on the population distribution, the solution may switch between regimes several times as we progress upstream of x. Regime (a) will occur first when $\eta(0) > \frac{L}{e}k_j$ and it will prevail until $\eta(\bar{x}) = \frac{L}{e}k_j$ for some $\bar{x}$. Then, regime (b) will

[11] Note that if some commuters from x depart from their ramp at t, commuters at $[x + dx]$ suffer a maximum additional delay at $t + dt$ equal to $\frac{k_j}{q_{max}}dx$.

happen as long as $(N(x) - N(\bar{x})) < \frac{L}{e} k_j (x - \bar{x})$. A new transition to (a) will happen for the next $\tilde{x}$ such that $N(\tilde{x}) - N(\bar{x}) = \frac{L}{e} k_j (\tilde{x} - \bar{x})$. In this case, it is easy to prove that $\tilde{C}(\tilde{x}) = \frac{N(\tilde{x}) + k_j \tilde{x}}{q_{max}} \frac{eL}{e+L}$ always.

Taking in account all these, the cost curve for the finite storage case will then be given by the lower envelope of the curve $\frac{N(x)+k_j x}{q_{max}} \frac{eL}{e+L}$ with slope lower of equal to $\eta(x)e$ and always bounded by the maximum single origin cost $\frac{N_T}{q_{max}} \frac{eL}{e+L}$. Logically, we recover the *zero storage* and *infinite storage* solution when $k_j \to 0$ and ∞ respectively.

Full analytical expressions can be obtained for the particular case of a population distribution with decreasing $\eta(x)$ – e.g., $\eta(x) = \eta e^{-\beta x}, \beta > 0$. In this case, the cost curve is defined as

$$\tilde{C}(x) = \begin{cases} \frac{N(x)+k_j x}{q_{max}} \frac{eL}{e+L} & \text{if } 0 \leqslant x \leqslant \bar{x} \\ \tilde{C}(\bar{x}) + \frac{N(x)-N(\bar{x})}{q_{max}} e & \text{if } \bar{x} \leqslant x \leqslant \ x^* \\ \frac{N_T}{q_{max}} \frac{eL}{e+L} & \text{if } x^* \leqslant x \leqslant \ell_T \end{cases} \tag{4.11}$$

where $\bar{x}$ is such that $\eta(\bar{x}) = \frac{L}{e} k_j$ and x^* is such that $\tilde{C}(x^*) q_{max} + (N(x^*) - N(\bar{x}))e = N_T \frac{eL}{e+L}$. Equation (4.11) can be simplified further assuming that only one of the two first domains takes place in the same cost curve. If $\eta(0) < \frac{L}{e} k_j$, then $\bar{x} = 0$ and $\tilde{C}(x) = \frac{N(x)}{q_{max}} e$ in $[0, x^*]$. If $\eta(0) > \frac{L}{e} k_j$ this is not true but one can consider instead that $\tilde{C}(x) = \frac{N(x)+k_j}{q_{max}} \frac{eL}{e+L}$ in $[0, x^*]$ since the difference with the exact solution is generally very small.

A equivalent expression for increasing $\eta(x)$ – e.g., $\eta(x) = \eta e^{-\beta x}, \beta < 0$– is,

$$\tilde{C}(x) = \begin{cases} \frac{N(x)}{q_{max}} e & \text{if } 0 \leqslant x \leqslant \tilde{x} \\ \frac{N(x)+k_j x}{q_{max}} \frac{eL}{e+L} & \text{if } \tilde{x} \leqslant x \leqslant x^* \\ \frac{N_T}{q_{max}} \frac{eL}{e+L} & \text{if } x^* \leqslant x \leqslant \ell_T \end{cases} \tag{4.12}$$

where $\bar{x}$ is such that $N(\bar{x})e = (N(\bar{x}) + k_j\bar{x})\frac{eL}{e+L}$ ($\bar{x} = 0$ if $\eta(0) > \frac{L}{e}k_j$) and x^* is such that $(N(\bar{x}) + k_j x)\frac{eL}{e+L} = N_T\frac{eL}{e+L}$.

4.3.3 Partial ramp priority approximations

For more general problems where $\alpha(x) = \alpha < \infty$, an exact closed-form expression $\tilde{C}(x)$ cannot be obtained (or when it is possible, it has a cumbersome expression). A reasonable approximation can be obtained instead by adopting the cost expressions of the full priority case, but assuming a distribution of population $\eta_\alpha(x) = \eta(x) + \frac{1}{\alpha}\frac{\partial\eta(x)}{\partial x}$, i.e., a cumulative population $N_\alpha(x) = N(x) + \frac{\eta(x)}{\alpha}$.[12]

To see why this approximation is reasonable consider the solution with partial ramp priority and zero freeway storage. Users at x take as given the decisions of downstream users and observe a time-dependent available capacity at the destination, $y(x,t)$. Unlike in the full ramp priority case, users at x do not block the freeway completely, but they always use a share α of this available capacity. Since, no freeway delays are possible, the solution interval $\Pi(x) = [t_s(x), t_f(x)]$ of arrivals for each location x will be given by the single bottleneck solution with capacity $\alpha y(x,t)$; i.e.,

$$\eta(x) = \int_{\Pi(x)} \alpha y(x,t)dt. \tag{4.13}$$

with $\tilde{C}(x) = -et_s(x) = Lt_f(x)$. To see then how cost changes as a function of $\eta(x)$, one can apply Leibniz's rule into (4.13) to see how $\Pi(x)$ changes; hence,

$$\frac{\partial\eta(x)}{\partial x} = \int_{\Pi(x)} \alpha\frac{\partial y(x,t)}{\partial x}dt + \frac{\mathrm{d}t_f(x)}{\mathrm{d}x}\alpha y(x,t_f(x)) - \frac{\mathrm{d}t_s(x)}{\mathrm{d}x}\alpha y(x,t_s(x)). \tag{4.14}$$

[12]Note that $\eta_\alpha(x)$ may be < 0. In this case, $\frac{\mathrm{d}\tilde{C}(x)}{\mathrm{d}x} < 0$.

Then, taking in account that $\frac{\partial y(x,t)}{\partial x} = -\alpha y(x,t)$ and $\mathrm{d}\tilde{C}(x) = -e\mathrm{d}t_s(x) = L\mathrm{d}t_f(x)$, equation (4.14) yields

$$\frac{1}{\alpha}\frac{\partial \eta(x)}{\partial x} + \eta(x) = \frac{\mathrm{d}\tilde{C}(x)}{\mathrm{d}x}\left(\frac{y(x,t_s(x))}{e} + \frac{y(x,t_f(x))}{L}\right). \tag{4.15}$$

If $\eta_\alpha = \frac{1}{\alpha}\frac{\partial \eta(x)}{\partial x} + \eta(x) > 0$, i.e., if $\Pi(x)$ always grows, then $y(x,t_s(x)) = y(x,t_f(x)) = q_{max}$ and

$$\frac{\mathrm{d}\tilde{C}(x)}{\mathrm{d}x} = \frac{1}{q_{max}}\left(\frac{1}{\alpha}\frac{\partial \eta(x)}{\partial x} + \eta(x)\right)\frac{eL}{e+L}. \tag{4.16}$$

Hence, the approximation is exact for solutions with zero-storage and $\eta_\alpha > 0$.

In the other cases, this approximation will work well as long $\eta_\alpha \geq 0$, i.e., as long as $\tilde{C}(x)$ does not decrease. An example is shown in Figure 4.3. The proposed closed-form approximation is compared to the exact solution, calculated using a continuous version of the algorithm in §3.4.2, for different storage capacity assumptions. Solutions correspond to the case of a homogenous distributed population $\eta = 1, \beta = 0$ and $\alpha(x) = 3.0$ (i.e, the continuous limit case of the discrete solutions in Figure 3.6). In all cases, the approximation preserves the right qualitative behavior.

A general expression of congestion cost for a decreasing density distribution $\eta(x)$ is then given by,

$$\tilde{C}_\alpha(x) = \begin{cases} (N(x) + \frac{\eta(x)}{\alpha} + k_j x)\frac{eL}{e+L} & \text{if } 0 \leqslant x \leqslant \bar{x} \\ \tilde{C}_\alpha(\bar{x}) + (N(x) - N(\bar{x}) + \frac{\eta(x)-\eta(\bar{x})}{\alpha})e & \text{if } \bar{x} \leqslant x \leqslant\ x^* \\ \frac{eL}{e+L}N_T & \text{if } x^* \leqslant x \leqslant \ell_T \end{cases} \tag{4.17}$$

with $\bar{x}$ s.t. $\eta(\bar{x}) + \frac{\eta(\bar{x})}{\tilde{\alpha}(\bar{x})} = \frac{L}{e}k_j$ and x^* s.t. $\tilde{C}_\alpha(\bar{x})q_{max} + (N(x^*) - N(\bar{x}) + \frac{\eta(x^*)-\eta(\bar{x})}{\alpha})e = N_T\frac{eL}{e+L}$.

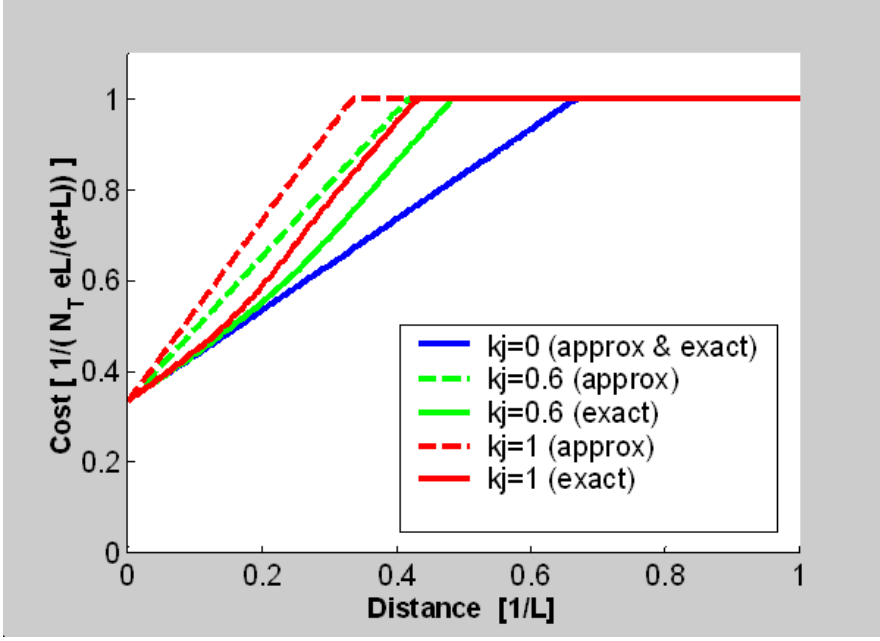

Figure 4.3. Continuous solution vs. closed-form approximation ($\eta(x) = \eta;\ \alpha(x) = 3$).

4.3.4 Final remarks

Expressions (4.11), (4.12) and (4.17) are remarkably simple and intuitive. From a qualitative point of view, they reveal the dependence of cost on queue spillovers (as discussed in §3.5). Note that the three domains correspond to different spillover situations and, clearly, increased storage capacity contributes to larger system costs. But, more importantly, they provide a inmmediate quantification of these effects and a clear measure of the mutual congestion interactions among populations located at different places. Commuting cost is shown to be mainly a function of the cumulative *intervening population* (i.e., the population located between the commuters residential location and its destination). Storage capacity available downstream contributes as additional intervening population. This should be intuitive since the storage capacity provides an opportunity to upstream commuters to queue in front of population in

x and hence, 'affect' the latter as if they were downstream. On the other hand, priority effects are measured as a function of the density/priority ratio (if density increases more rapidly than priority, cost necessarily must increases with distance and viceversa). One can also observe that there is a threshold level for k_j beyond which increases in k_j leave unaffected the cost curve. (The first domain in (4.11) and (4.17) never arises for a sufficiently large k_j and the cost in other domains is not affected by k_j.) This is interesting since it suggests that there is a sprawl threshold beyond which the system congestion level stagnates.[13]

All these causal relationships differ greatly from the implicit assumptions made in the traditional congestion models. Hence, the analysis done here clearly opens the door to reconsider most of the analysis of urban location based on congestion as a promising new line of research.

[13]Recall from §3.5 that $\hat{k}_j = \frac{k_j}{N_T \ell_T}$ was an indirect measure of sprawl.

Chapter 5

Some Generalized Models of Departure Time Equilibrium

EVERY MODEL abstracts from reality in order to focus on the essentials and avoid the details. In that regard, some simplifying assumptions were adopted in the previous chapters, mainly that the freeway network was homogeneous and that all commuters had the same deadline. Under these assumptions, the morning commute equilibrium problem could be treated analytically and general qualitative insights were gained.

This chapter extends the equilibrium analysis to more general instances. The goal is twofold. First, we seek to show that the main insights derived in chapters 2 and 3 about the spatial behavior of congestion hold under more general assumptions. Secondly, we want to provide some additional technical analysis that can be used in the future to define general algorithms to solve the dynamic traffic assignment problem under the KW model of traffic flow. For that purpose, we first extend the analysis of the two-origin network in chapter 2: in section 5.1, the case of a heterogeneous network (i.e., links are not identical) and, in section 5.2, the case where commuters

have different preferred arrival times. Then, in section 5.3, the freeway problem is briefly revisited to incorporate these assumptions. Finally, section 5.4 briefly discusses possible extensions to problems with route choice and multiple destinations.

5.1 Two-origin network: Heterogeneous links

We shall consider a network where the upstream links AM and BM have capacities q_{AM} and q_{BM} smaller than the capacity of the downstream link MD, q_{MD}; see Figure 5.1.[1] We shall first explore the traffic dynamics and briefly summarize how solutions are obtained. Then, some equilibrium solutions examples are discussed. As in chapter 2, we present solutions for: (i) cases with no bottleneck restriction at D ($q_D(t) = q_{MD}$) where queues cannot form on link MD and merging effects dominate, and (ii) cases with (time-dependent) flow restrictions at D ($q_D(t) \leqslant q_{MD}$) where queue spillovers affect performance.

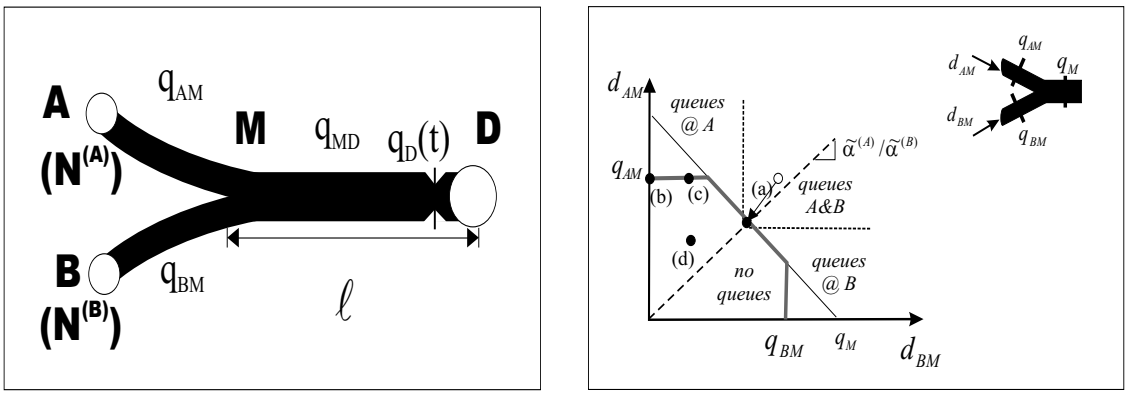

Figure 5.1. Heterogeneous 2-origin network.

[1] When the upstream capacities are larger (i.e., $q_{AM}, q_{BM} > q_{MD}$) the outflow is always controlled by the downstream capacity q_{MD} and hence, the overall behavior of the system is as if all links had equal capacity.

5.1.1 General solution procedure

Restricted capacity upstream may cause bottleneck starvation; i.e., the capacity available at M or D may not be fully used even if a queue exists in one of the upstream approaches. As result, the equilibrium problem may not always be treated as a single bottleneck in the aggregate or be decomposed by origin. It is still possible, however, to obtain the equilibrium solution as a function of the times $\{t_s^{(r)}, r = A, B\}$ when commuters in each origin first experience any delay. If these are known, so are the equilibrium delays $\{\tau^{(r)}, r = A, B\}$; see Property UE(1) in §2.2.3. Hence, we shall show how the traffic solution for each origin can be constructed as a function of the $\{\tau^{(r)}, r = A, B\}$. The basic idea is that one can express all traffic information as a function of the arrival time to the destination t and construct incrementally the solution advancing over t. That is, one can build the cumulative curves in Figure 2.4 by staking successively horizontal slices on top of each other.[2]

We shall then express first link and merge behavior as a function of the delay information. Note that link and merge behavior follow the KW and merge rules outlined in §3.2 for general networks.

Link MD behavior. First, the departure rate at D, d_D ($\equiv d_{MD}$, using the link notation in chapter 3) will be equal to q_{MD} if a queue exists at link MD; otherwise, it will be equal to the observed departure rate at M, $d_M(\equiv a_{MD})$; i.e.,

$$d_D(t) = \begin{cases} q_{MD} & \text{if } \tau_{MD}(t) > 0 \text{ (i.e., queues in link } MD) \\ d_M(t) & \text{if } \tau_{MD}(t) = 0 \text{ (i.e., no queues in link } MD) \end{cases} \tag{5.1}$$

On the other hand, the available capacity downstream of M, $d_M^{max}(t_M) \leqslant q_{MD}$ will

[2]This decomposition by arrival time has been applied to route choice equilibrium in Akamatsu and Kuwahara [1999]; Kuwahara and Akamatsu [1993].

depend on spillovers happening or not, i.e.,

$$d_M^{max}(t_M) = \begin{cases} d_D^D(t_M) \equiv d_D(t_M - \frac{\ell}{\hat{w}}) & \text{if } \tau_{MD}(t) = \nu_{MD}(t) \text{ (i.e., spillover)} \\ q_{MD} & \text{if } \tau_{MD}(t) < \nu_{MD}(t) \text{ (i.e., no spillover)} \end{cases} \tag{5.2}$$

where $t_M = t - \tau_{MD}(t)$ is the time of pass through point M of a vehicle arriving at the destination at time t.

Merge behavior. The departure rate at M depends on the merge behavior, which is controlled by the downstream available capacity $d_M^{\max}$ and the maximum departure rates from each approach $d_{rM}^{\max}$; see equations (3.8)-(3.9). Figure 5.1b shows the merge diagram when the approaches have reduced capacity (this figure is equivalent to Figure 3.4).

The maximum departure rate from each approach $d_{rM}^{\max}$ depends on the queuing status upstream, which is given by the difference between the equilibrium delays $\tau^{(r)}$ and τ_{MD}. If $\tau^{(r)} > \tau_{MD}$, a queue exists upstream of approach rM. If $\tau^{(r)} = \tau_{MD}$ queues are not present and outflow at the upstream approach must be equal to the origin arrival rate (that is, the departure rate from the origin). If $\tau^{(r)} < \tau_{MD}$, commuters at origin r will not flow during at this time since the common delay at link MD is higher than the equilibrium delay. Hence,

$$d_{rM}^{\max}(t_M) = \begin{cases} q_{rM} & \text{if } \tau^{(r)}(t) > \tau_{MD}(t) \text{ (i.e., queues)} \\ a^{(r)}(t_M) & \text{if } \tau^{(r)}(t) = \tau_{MD}(t) \text{ (i.e., no queue, feasible)} \\ 0 & \text{if } \tau^{(r)}(t) > \tau_{MD}(t) \text{ (i.e., no queue, non-feasible)} \end{cases} \tag{5.3}$$

Then, four sustained scenarios are only possible in agreement with (3.9), assuming

from Property UE(2) that $\tau^{(A)} \geqslant \tau^{(B)}$:[3]

(a) $\min\{\tau^{(A)}(t), \tau^{(B)}(t)\} > \tau_{MD}(t)$ (i.e., queues in both approaches). Then, $d_M(t_M) = d_M^{\max}(t_M)$ and vehicles arrive at the ratio $\alpha_D^{(A)}(t)/\alpha_D^{(B)}(t) = \tilde{\alpha}^{(A)}/\tilde{\alpha}^{(B)}$.

(b) $\tau^{(A)}(t) > \tau_{MD}(t) > \tau^{(B)}(t)$ (i.e., a queue in one approach and no departures in the other). Then, $d_M(t_M) = \min\{q_{AM}, d_M^{max}(t_M)\}$.

(c) $\tau^{(A)}(t) > \tau_{MD}(t) = \tau^{(B)}(t)$ (i.e., a queue in one approach and some departures in the other). Then, $d_M(t_M) = q_{AM} + a^{(B)}(t_M)$ where $a^{(B)}(t_M) = d^{(B)}(t)/(1+\dot{p}(t))$. Hence, the proportion of flows is $\alpha_D^{(A)}(t)/\alpha_D^{(B)}(t) = q_{AM}/a^{(B)}(t_M)$. This scenario is sustained if $a^{(B)}(t_M) + q_{AM} < d_M^{max}(t_M)$.[4]

(d) $\tau^{(A)}(t) = \tau^{(B)}(t) = \tau_{MD}(t)$ (i.e., no queues in any approach). Then, $d_M(t) = a^{(A)}(t) + a^{(B)}(t)$. Furthermore, any split of flows between A and B is feasible since the equilibrium arrival curves must satisfy $d_M(t) = a^{(A)}(t_M) + a^{(B)}(t_M) = d^{(A)}(t)/(1 - \dot{\tau}^{(A)}) + d^{(B)}(t)/(1 - \dot{\tau}^{(B)}) = (d^{(A)}(t) + d^{(B)}(t))(1 + \dot{p}(t))$.[5] This scenario is sustained as long as $\tau^{(A)} = \tau^{(B)} \leqslant \nu_{MD}$.

The four situations are summarized in Figure 5.1b.

The full traffic solution can be constructed sequentially as a function of t for given equilibrium delay curves $\{\tau^{(r)}, r = A, B\}$ in the following manner. For each $t' < t$, we know: the history of departures from D for each origin $\{D_D^{(A)}(t'), D_D^{(B)}(t')\}$, the actual delay experienced in link MD for all vehicles arrived before t $\{\tau_{MD}(t')\}$, the time of pass through the merge M, $t_M = t - \tau_{MD}$, and the spillover delay $\nu_{MD}(t)$ – the horizontal difference of $D_D(t)$ and the shifted $D_D(t - \ell/\hat{w}) + k_j\ell$. Then, we can use

[3] Symmetric results apply for $\tau^{(A)} \leqslant \tau^{(B)}$.

[4] This situation never arose on the homogeneous network since $q_{AM} = q_{MD} \geq d_M^{max}(t_M)$.

[5] As shown later, if commuters have different deadlines the indetermination can be avoided.

equations (5.1)-(5.3) and the four cases above to determine directly the increments with t $\mathrm{d}D_D(t)$ and $\mathrm{d}D_M(t_M)$ and $\{\alpha_D^{(A)}(t), \alpha_D^{(B)}(t)\}$. Hence, we can also obtain the increments in $\mathrm{d}\tau_{MD}(t)$ and $\mathrm{d}\nu_{MD}(t)$ and construct the solution incrementally.

The equilibrium $\{t_s^{(r)}, r = A, B\}$ – those that yield origin-specific departure processes $\{D_D^{(A)}, D_D^{(B)}\}$ with total outflows equal to the populations $\{N^{(A)}, N^{(B)}\}$ – must be obtained by iteration. This is a two-dimensional search which, however, is straightforward since each origin total outflow is increasing with earlier $t_s^{(r)}$. Furthermore, in many cases the search can be decoupled by origin as in the homogenous case, but this can not always guaranteed.

5.1.2 No downstream restrictions

As in the homogeneous case, when no restrictions exist downstream of merge M, the system behaves as a pair of single-origin bottlenecks with departure rates coupled by the merging rule and, in this case, the solution can be decoupled by origin. Figure 5.2 illustrates the combined final solution. Commuters from one origin (B in the figure) flow through the bottleneck during an interval when the other approach is always queued. Therefore, they use a fixed share of the capacity $\alpha_M^{(B)}(t) = \tilde{\alpha}^{(B)}$, i.e., case (a) above. Hence, the equilibrium solution for this origin (i.e., the interval $[t_s^{(B)}, t_f^{(B)}]$) is obtained by solving a single bottleneck equilibrium with capacity $\tilde{\alpha}^{(B)} q_{MD}$. Origin A-commuters flow at $\tilde{\alpha}^{(A)} q_{MD}$ during the interval $[t_s^{(B)}, t_f^{(B)}]$ and at capacity q_{AM} otherwise. We obtain the arrival and departure curve solving a time-dependent capacity single bottleneck.

From the figure, it is again evident that B-users experience less cost than A-users if $N^{(B)}/\tilde{\alpha}^{(B)} < N^{(A)}/\tilde{\alpha}^{(A)}$. The opposite is true if $N^{(B)}/\tilde{\alpha}^{(B)} > N^{(A)}/\tilde{\alpha}^{(A)}$. In the limit case, $N^{(B)}/\tilde{\alpha}^{(B)} = N^{(A)}/\tilde{\alpha}^{(A)}$, the cost coincide for both origins (and the solution is as

in the homogeneous case, since the merge section is always at the full capacity $q_M D$). Therefore, we recover the origin order by cost based on the *population-to-priority* ratio.

As in the homogeneous case, the performance of the system can be improved by manipulating the priorities $\{\tilde{\alpha}^{(A)}, \tilde{\alpha}^{(B)}\}$. Note, however, that increasing the priority to origin B decreases the cost for commuters on this origin but increases the cost for the commuters in A. Therefore, there is a trade-off on the possible cost-improving ramp metering strategies. As show in appendix A, however, it is still optimal to give as much priority to the one the approaches as possible.

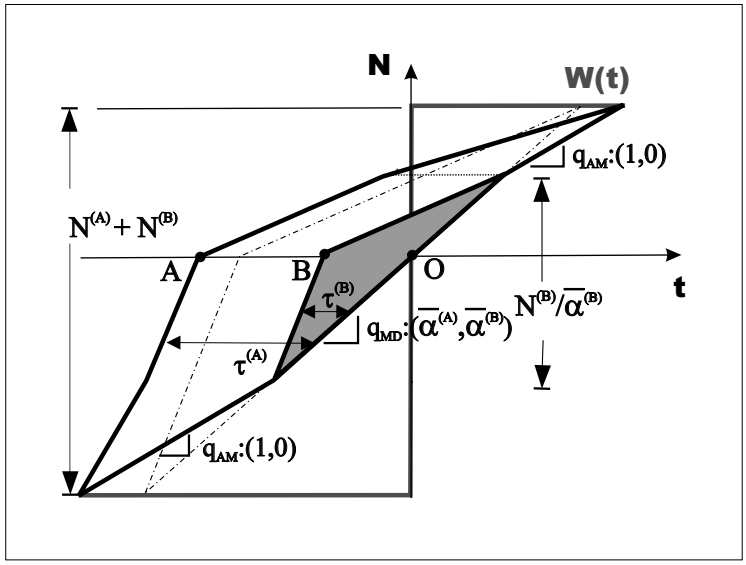

Figure 5.2. Equilibrium solution: heterogeneous links, no queues in link MD.

5.1.3 Downstream restrictions

The equilibrium patterns when queues happen on link MD can be very diverse, depending on the characteristics of the restriction at D and the values of the upstream capacities. Figure 5.3 shows equilibrium solutions for the case where a perennial

bottleneck with capacity $q_D(t) = q_D \leqslant q_{MD}$ exists at D. We have assumed the origin population proportions such that the cost for B-users is always smaller; i.e., the equivalent to solution type 1B in chapter 2 when $t_s^{(B)} < t_s^{(A)}$. Each solution corresponds to a different assumption about the upstream capacities. The different queuing episodes and the corresponding origin-specific outflow shares are defined in agreement with cases (a)-(d) in §5.1.1: shaded areas correspond to intervals where queues exists upstream of both approaches and the merge is saturated, i.e., case (a); non-shaded areas correspond to intervals where only A-commuter flows, i.e., case (b); cross-hatched areas to intervals where queues exists upstream of A but B-commuters use the portion of downstream merge capacity not used by A-users, i.e., case (c).[6]

From Figure 5.3, it is apparent that the same overall qualitative behavior observed on the homogeneous network holds now. Fundamentally, the performance of the system worsens as the length of the common link MD increases (i.e., we shift away the thick dashed curve in Figure 5.3). Cost for origin B increases while the cost for origin A remains unchanged. This comes at no surprise since, as explained in chapter 2, the common link gives an opportunity for A-users to mingle in a common queue with B-users and this is independent of the upstream capacities values. Furthermore, the total cost may also increase as we increase the capacity of the upstream link AM (total cost always increases if $q_{AM} > q_D$ and may increase in some cases when $q_{AM} < q_D$; see Figure 5.3b-c). This is yet another example that a 'capacity-increasing paradox' can arise in dynamic scenarios as a consequence of commuter departure time adaptation only. This was already pointed out in [Arnott et al., 1993a] using a model with point queues. Our analysis shows, however, that such paradoxes are less prevalent than stated there since the influence of an increase in the capacity of an

[6]Case (d) only arises in solutions where $t_s^{(B)} = t_s^{(A)}$.

upstream link is somewhat limited by the merging interactions, which are neglected in that latter work. A more detailed discussion can be found in appendix A.

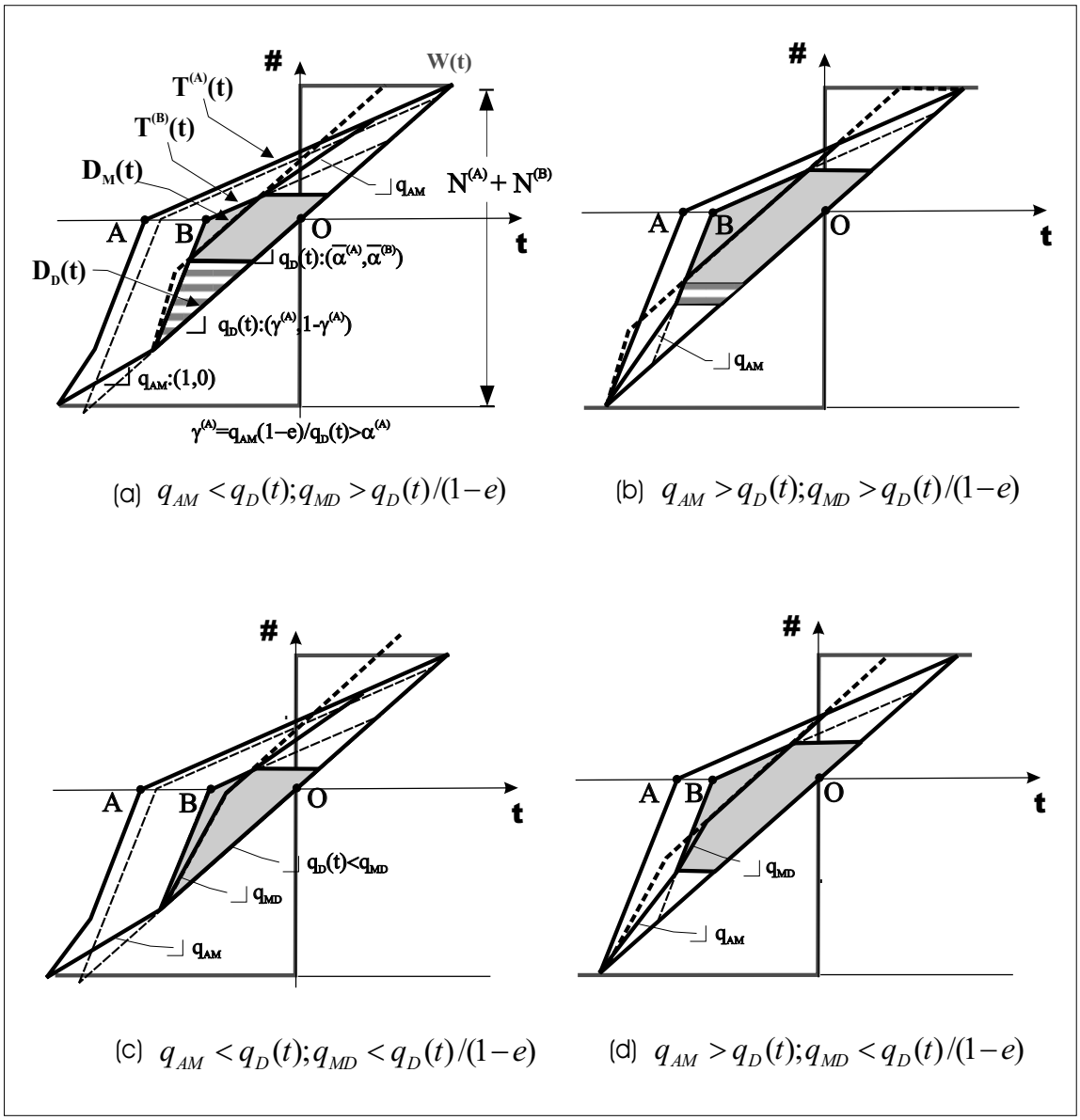

Figure 5.3. Equilibrium solution: heterogeneous links, queues in link MD.

From an algorithmic point of view, note that when $t_s^{(B)} \to t_s^{(A)}$, we recover the homogeneous case solution; i.e., in this case, the solution is not affected by

the upstream capacities since queues in both approaches happen simultaneously and the merge works always at capacity during these intervals. This is fortunate because we can use the same criteria to discriminate among the different types of solutions that arise with different populations: B-commuter cost is lower when $0 < N^{(B)}/\tilde{\alpha}^{(B)} \leqslant (N^{(A)}+N^{(B)})-N_U$, and A-commuter cost is lower when $N^{(A)}/\tilde{\alpha}^{(A)} \leqslant (N^{(A)} + N^{(B)}) - N_U)$ where N_U is the total number of vehicles flowing through the merge in non-congested conditions assuming the homogenous solution. This criterion is independent of the values of the capacities upstream.

5.2 Two-origin network: Different deadlines

Further realism is added if we consider that commuters have different desired times of arrival to the destination. The distribution of *deadlines* of commuters in each origin can be conveniently represented by a cumulative curve $W^{(r)}(t)$ representing the number of commuter wishing to arrive to D before time t (where $W^{(r)}(T) = N^{(r)}$). $W^{(r)}(t)$ is commonly assumed to be S-shaped, such that a major intensity of desired arrivals occurs at the peak of the rush hour and a single main global congestion period happens [Smith, 1984] – this is not strictly necessary, however. We shall further consider that the distribution of deadlines is homogeneous across origins, i.e., $W^{(r)}(t) = \eta^{(r)}\ W(t)\ \forall t$. This condition is not unreasonable and it is very convenient to extend the single deadline rationale naturally. Note, for instance, that $\eta^{(r)}$ plays now the role of population $N^{(r)}$, i.e., the *population-to-priority* ratio is defined as $\eta^{(r)}/\tilde{\alpha}^{(r)}$. The implication of considering deadlines distribution non-homogeneous of across origins are also briefly explored at the end of section 5.2.2.

5.2.1 Equilibrium under different deadlines

The equilibrium conditions must take in account that the commuters decisions and the equilibrium cost depends on commuter's deadline. It can be shown that there is always an equilibrium pattern in which commuters in each origin arrive to the destination in the order of desired deadlines; see Daganzo [1985] for a formal proof with a single origin; Kuwahara [1990] for several origins. With this condition (that we call fist-desired-first-in or *FDFI*) one can always define for each arrival time a unique schedule delay $s^{(r)}(t)$ such that $D^{(r)}(t) = W(t - s^{(r)}(t))$. Hence,

$$\dot{s}^{(r)}(t) = 1 - d^{(r)}(t)/w^{(r)}(t_w) \tag{5.4}$$

with $t_w = t - s^{(r)}(t)$ being the deadline of the commuter arriving at time t and $d^{(r)}$ and $w^{(r)}$ the respective derivatives of $D^{(r)}$ and $W^{(r)}$. Furthermore, the cost for origin r commuters can then be uniquely expressed as $C^{(r)}(t) = \tau^{(r)}(t) + p(s^{(r)}(t))$. Taking derivatives with respect to arrival time, we obtain the following necessary condition for equilibrium when equilibrium delays are positive and origin outflows continuous:

$$\partial C^{(r)}/\partial t = 0 \Rightarrow \dot{\tau}^{(r)}(t) = -\dot{p}(s^{(r)}(t)) \quad \forall \tau^{(r)}(t) > 0. \tag{5.5}$$

Similarly to the case with a single deadline, we could define for each origin the time at which the first commuter suffers any delay, $t_s^{(r)}$. For each set of $\{t_s^{(r)}, r = A, B\}$, equation (5.4) and (5.5) could be used in combination with the traffic model rules in §5.1.1 to define incrementally $\{\tau^{(r)}(t), s^{(r)}(t), D^{(r)}(t)\}$ for each origin as function of the departure time from D. Equilibrium would be reached by finding $\{t_s^{(r)}, r = A, B\}$ that make the equilibrium delay $\tau^{(r)}(t)$ and scheduled delay $s^{(r)}(t)$ vanish at the same departure time $t_f^{(r)}$.

We shall present typical solutions with a homogeneous network (with link capacity q_{max}) and discuss qualitatively some of the implications of the results.

5.2.2 No downstream restrictions

We first consider the case where no queues form at link MD, $q_D(t) = q_D = q_{max}$. Figure 5.4 illustrates the final solution – we assume that the queuing period at one approach (i.e., B in the figure) always happens when the other approach is congested, i.e., $\Pi^{(B)} = [t_s^{(B)}, t_f^{(B)}] \subset [t_s^{(A)}, t_f^{(A)}] = \Pi^{(A)}$. The equilibrium solution for this origin is obtained by solving a single bottleneck equilibrium with capacity $\tilde{\alpha}^{(B)} q_D$ and deadline curve $W^{(B)}(t) \equiv \eta^{(B)} W(t)$. Note that outside the interval $[t_s^{(B)}, t_f^{(B)}]$ B-commuters still flow at a rate $d^{(B)}(t) = w^{(B)}(t)$ since they must suffer no delay or schedule delay. More importantly if $\eta^{(B)}$ is small enough no queues appear in this approach and people experience no congestion costs at all. Origin A-commuters must flow at $\tilde{\alpha}^{(A)} q_D$ during the interval $[t_s^{(B)}, t_f^{(B)}]$ and at $q - w^{(B)}(t)$ otherwise. We obtain the arrival and departure curve solving a time-dependent capacity single bottleneck with deadline $W^{(A)}(t) \equiv \eta^{(A)} W(t)$.

The individual solutions for each origin are presented in Figure 5.4a. As in chapter 2, we rescale and superimpose the two solutions in a unique aggregated in Figure 5.4b. Note that $T^{(A)}$ and $T^{(B)}$ are *A/D curves* which define the equilibrium delays for each origin.At the same time, $\tilde{W}^{(A)}(t)$ and $\tilde{W}^{(B)}(t)$ are *re-scaled deadline-schedule curves* (or D/W curves) that define the scheduled delays as a function of the arrival time. Although the slopes of the curves $T^{(A)}$ and $T^{(B)}$ must be the same for both origins since commuters share a common penalty function, the departure times from D corresponding to the vehicles that arrive on time, which are actually those that experience the highest cost in each origin, do not necessarily coincide for both origins.

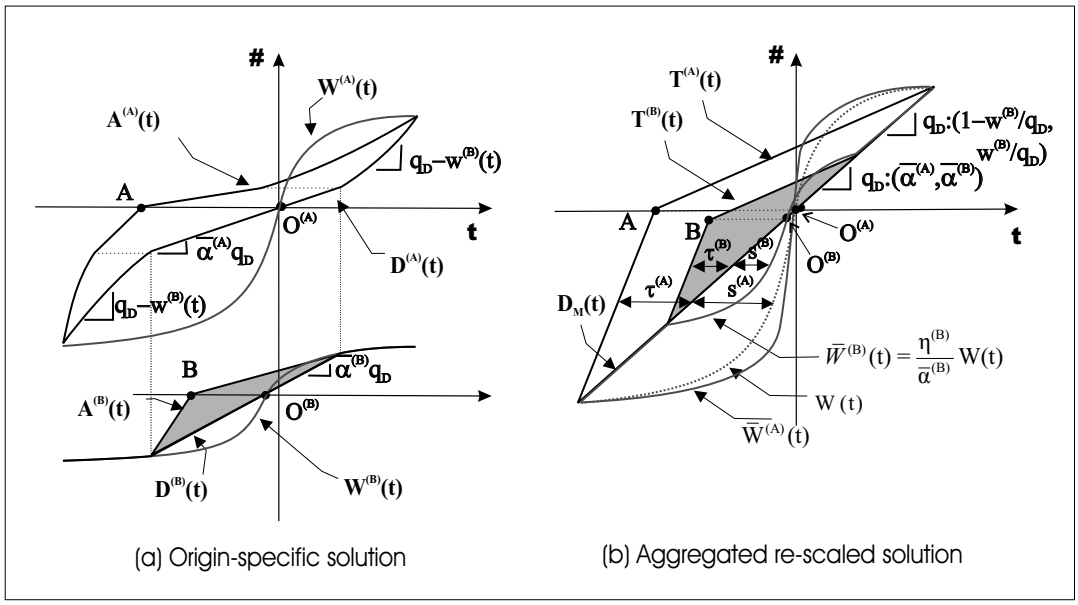

Figure 5.4. Equilibrium solution: different deadlines, no queues in link MD.

Note that the aggregated departure curve from M, D_M, necessarily coincides with the departure curve of a single bottleneck solution with capacity q_{max} and aggregated deadline curve $W(t) = W^{(A)}(t) + W^{(B)}(t)$. Furthermore, since $t_s^{(A)} = t_s$ and $t_f^{(A)} = t_f$, the delays for origin A must also coincide with those of the single origin solution (the schedule delay $s^{(A)}(t)$ may not coincide, however). Whenever approach B is queued, $\tilde{W}^{(B)}(t)$ must increase at a rate $\eta^{(B)} w(t)/\tilde{\alpha}^{(A)}$. Hence, as long as $\eta^{(B)}/\tilde{\alpha}^{(B)} < 1$ (i.e., $\eta^{(B)}/\tilde{\alpha}^{(B)} < \eta^{(A)}/\tilde{\alpha}^{(A)}$) B users queue always when the other approach is congested as we assumed; i.e., $t_s^{(B)} > t_f^{(A)}, t_f^{(B)} < t_f^{(A)}$). In this case, the cost associated with each deadline is always higher for A-vehicles. When $\eta^{(B)}/\tilde{\alpha}^{(B)} = \eta^{(A)}/\tilde{\alpha}^{(A)}$ solutions for both origins overlap (i.e., $t_s^{(B)} = t_s^{(A)}, t_f^{(B)} = t_f^{(A)}$) and the cost associated with each deadline is equal for both origins.

This result holds whenever the distribution of deadlines is assumed homogenous across origins,[7] and is important, because implies that we can establish a ordering

[7]The proof seems relatively straightforward from our analysis

of origin delays as a function of the ordering in cost $C^{(r)}$ (or $t_s^{(r)}$); i.e., Property UE(2) in chapters 2 and 3 still holds. This is relevant to extend the different deadline analysis to multi-origin networks. For the sake of completeness, Figure 5.5 shows a typical solution with a distribution of deadlines *non-homogenous across origins*. An extreme case is represented where all B-commuters desire to arrive before time O and all A-commuters afterwards. In this case, the aggregated departure curve from D still coincides with that of a single origin. However, the ordering in delays does not necessary hold and the single origin delays represents now an upper bound to the maximum delays in any of the two origins.

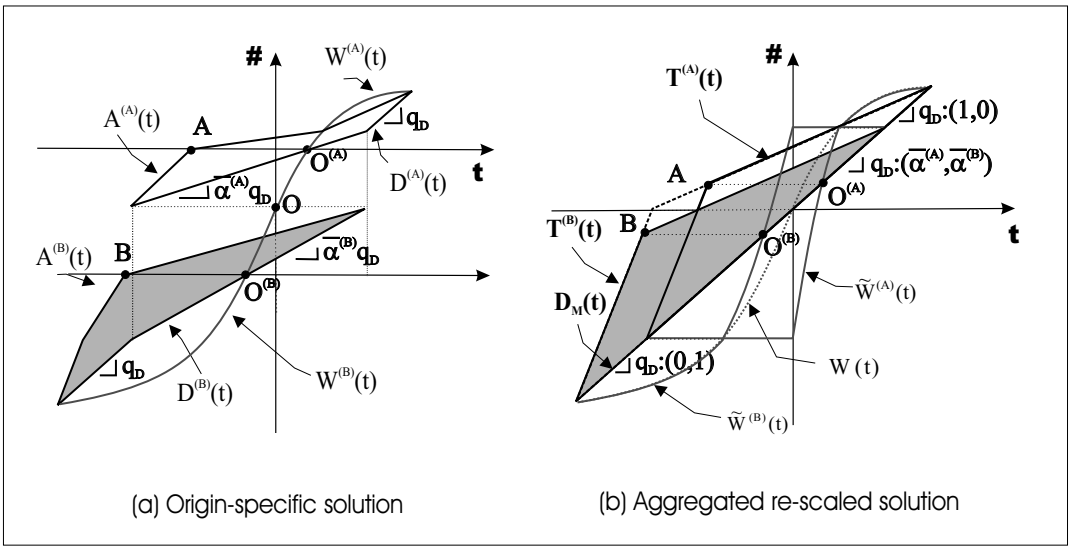

Figure 5.5. Equilibrium solution: non-homogeneous deadlines, no queues in link MD.

5.2.3 Downstream restrictions

Figure 5.6 show the aggregated final solution for the case of a permanent bottleneck at D and Figure 5.7 for the case of a time-dependent restriction at D. Different types of solution may arise depending on the values of $\eta^{(B)}/\tilde{\alpha}^{(B)}$ and $\eta^{(A)}/\tilde{\alpha}^{(A)}$. In

each figure, cases (a) to (d) correspond to different solutions for increasing $\eta^{(B)}/\tilde{\alpha}^{(B)}$ with $\eta^{(B)}/\tilde{\alpha}^{(B)} > \eta^{(A)}/\tilde{\alpha}^{(A)}$; i.e., a lower origin B *population-to-priority ratio*. Symmetric results follow otherwise. (Note that only the aggregated deadline curve W and the origin-B rescaled deadline curve $\tilde{W}^{(B)}$ are shown; $\tilde{W}^{(A)}$ is omitted since it does not provide additional information.)

To see that the solutions shown are equilibria consider the following:

- When $\tau^{(A)}(t) = \tau^{(B)}(t) = \tau_{MD}(t) \leqslant \nu_{MD}(t)$ (i.e., no queues upstream, all commuters suffer common delay at MD), the FDFI condition requires $\dot{s}^{(A)}(t) = \dot{s}^{(B)}(t) = \dot{s}(t)$; hence, the outflows from each origin must be in the same proportion as the desired arrivals $(\alpha_D^{(A)}, \alpha_D^{(B)}) = (\eta^{(A)}, \eta^{(B)})$. Graphically, we have that $\tilde{w}^{(B)}(t_w) = w(t_w) = \tilde{w}^{(A)}(t_w)$.

- When $\tau^{(A)}(t) > \tau^{(B)}(t) = \tau_{MD}(t)$ (i.e., B-commuters suffer no delay upstream of M), equilibrium requires that $s^{(B)}(t) = 0$ and $\dot{s}^{(B)}(t) = 0$. Therefore, $d^{(B)}(t) = w^{(B)}(t)$ or equivalently, $(\alpha_D^{(A)}, \alpha_D^{(B)}) = (1 - w^{(B)}/q_D(t), w^{(B)}/q_D(t))$. Graphically, we have that $\tilde{W}^{(B)}(t) = D_D(t)$. Note that this is only sustained if $\dot{\tau}_{MD}(t) = 0$.

- When $\tau^{(A)}(t) > \tau^{(B)}(t) > \tau_{MD}(t)$ (i.e., B-commuters suffer delay upstream of M), then $(\alpha_D^{(A)}, \alpha_D^{(B)}) = (\tilde{\alpha}^{(A)}, \tilde{\alpha}^{(B)})$. Hence $\tilde{W}^{(B)}(t_w) = \eta^{(B)} W(t_w)/\eta^{(B)}$ where $t_w = t - s^{(B)}(t)$.

The transitions between the different states correspond to the commuter in B that, given his deadline, is indifferent between departing in one or other state.

Type 1 solutions corresponds to where $\eta^{(B)}/\tilde{\alpha}^{(B)}$ is small enough such that most B-commuters experience less cost than corresponding A-commuters. Type 2 corresponds to situations where only commuters from B arriving late experience less cost than the

A commuters having the same deadline. Finally, type 3 corresponds to cases where A and B commuters having the same deadline experience the same cost. In the limit when $\eta^{(A)}/\tilde{\alpha}^{(A)} = \eta^{(B)}/\tilde{\alpha}^{(B)}$ we have a perfectly symmetric solution, as in the no spillover case. Note that unlike in the single deadline case, a unique flow assignment satisfying the *FDFI* solution exists for each value of $\eta^{(A)}/\tilde{\alpha}^{(A)}$ and $\eta^{(B)}/\tilde{\alpha}^{(B)}$.

We see that increasing the priority of the smaller population approach always reduces cost for this population. At the same time, increasing link MD length or its storage capacity increases the total cost for this population. Note however that the cost savings do not accrue to the complete approach population equally but depend on commuter's deadline. This is relevant since the average cost may not change across origins as much as in the single deadline case, suggesting that the spatial effects of queues are less important when commuters have different deadlines.

From an algorithmic point of view, note that since Property UE(2) holds for the case with homogeneous distribution of deadlines across origins, the solution for the homogeneous network can be constructed following the rationale of the single deadline case. First, the aggregated departure curve $D_D(t)$ and the maximum delays $\tau^M(t)$ are obtained as if we had a single bottleneck problem with deadline curve $W(t) = W^{(A)}(t) + W^{(B)}(t)$ and capacity $q_D(t) < q$ (Step 1). Then, the delays in link MD $\tau_{MD}(t)$ can be inferred as in chapter 2 as a function of $\tau^M(t)$ and the spillover delays $\nu_{MD}(t)$ given by Newell's Shift (Step 2). Then, we need to obtain the solution for the origin with smaller population-to-priority ratio as a function of $[t_s^{(B)}, t_f^{(B)}]$ (Step 3), where now $t_s^{(B)}$ and $t_f^{(B)}$ represent the first and last departure time from D at which a commuter from B experiences any delay upstream of M. The solution of this step is quite tedious however, since for each candidate $t_s^{(B)}$ it is necessary to identify the commuters for which the transition between different queuing states occur.

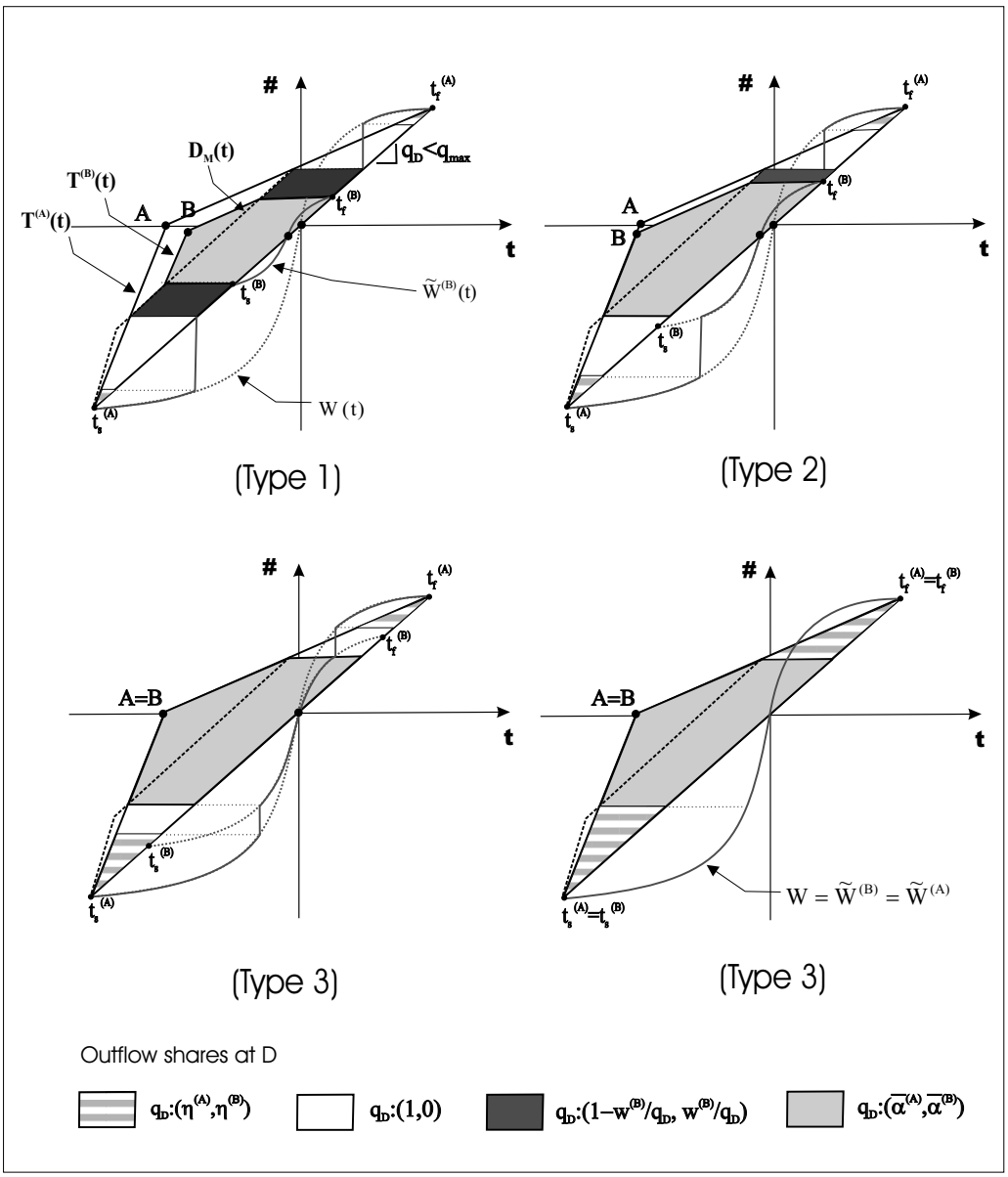

Figure 5.6. Equilibrium solutions: different deadlines, queues in link MD (Permanent bottleneck).

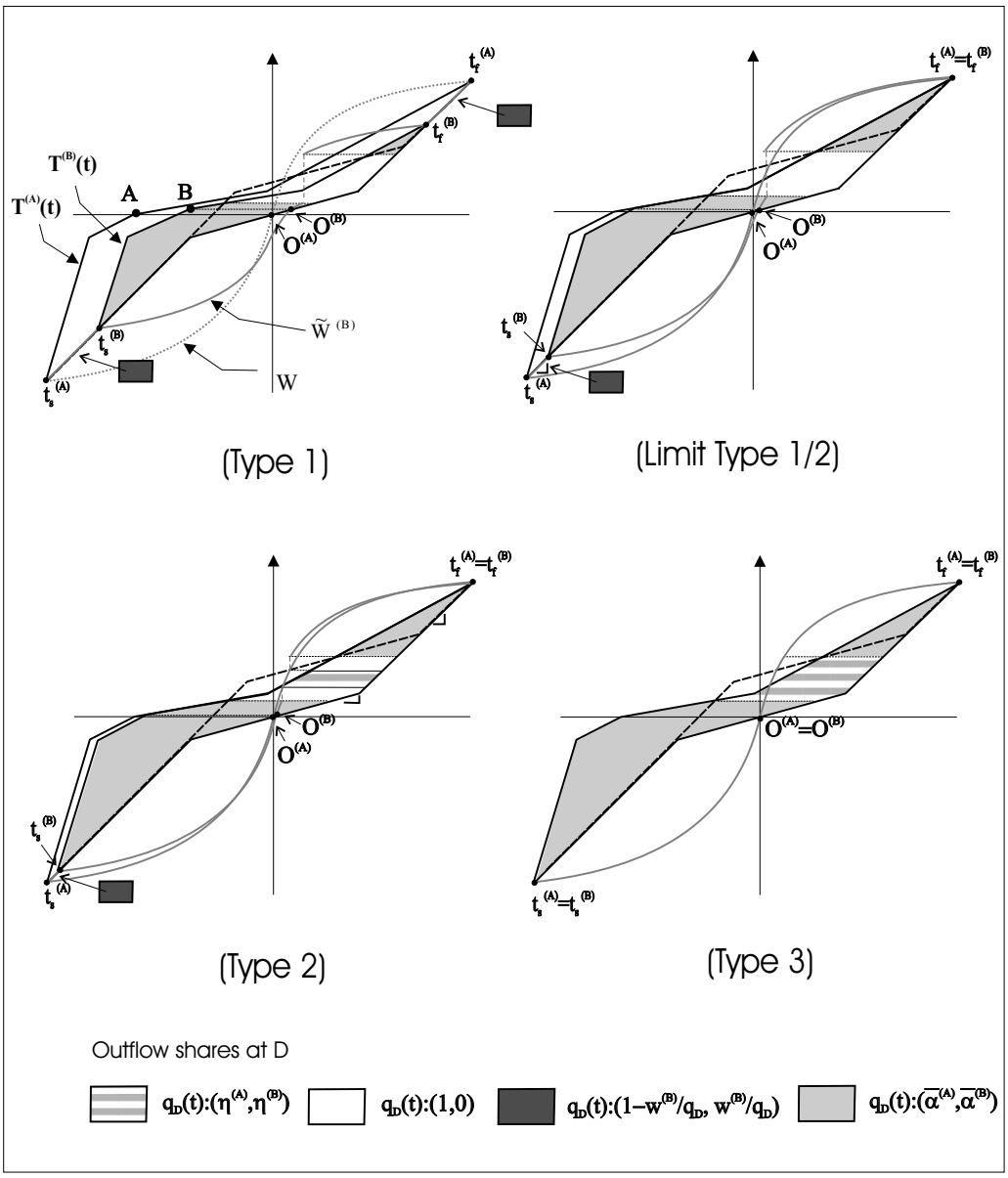

Figure 5.7. Equilibrium solutions: different deadlines, queues in link MD (Time-dependent bottleneck).

5.3 The freeway model revisited

The analysis in the previous sections can also be extended to the case of the linear freeway of chapter 3.

5.3.1 Heterogeneous freeway/tree networks

Reduced capacities at upstream links may restrict flow and affect the observed departure curve at any downstream node. Hence, Property UE(3) in chapter 3 does not strictly hold for a freeway with heterogeneous links and the equilibrium solution at any node may depend on the distribution of population across upstream origins.

One can still try to use the recursive procedure presented in §3.4. An equivalent two-origin problem where the upstream links have different capacity will be solved at each merge as indicated in §5.1.1. The actual departure curve upstream of a node will be updated according to the lower capacity of the upstream approach. Since Property UE(3) does not strictly hold, this procedure may not always yield a feasible equilibrium but it is reasonable to believe that the solution obtained will likely differ very little from the true equilibrium and provide a sufficiently accurate measure of the variation of origin costs. Note, for example, that for the two origin network the solution for commuters in the lower cost origin is only affected by the fact that the high-cost commuters will tend to use the maximum capacity available upstream of the merge during the times when the former commuters flow; it does not depend, however, on the actual solution of the high cost origin, since delays for this origin are always higher. Hence, if the distribution of populations on the freeway network are such that costs at origins upstream of node i is always larger then the equilibrium cost at origin i (i.e., $\Pi^{(i)} \subset \Pi^{(j)}\ \forall j > i$), the equilibrium at origin i can be calculated as if the whole upstream population came from a single origin and used the full capacity available

at the link upstream; i.e., the sequential procedure must yield exact solutions.

In regard to the change of congestion cost with distance, reduced capacity upstream will lower the cost of the downstream origins and increase those of upstream origins (as it is shown in the two-origin network solution). At the same time, with narrower links queues will spillover more easily. This suggests that stepper increasing gradients on the congestion cost with distance will be found in freeways with decreasing capacity upstream. Still, there will be a threshold location beyond which congestion costs stagnate. Further analysis needs to be pursued to provide full evidence, though.

5.3.2 Different deadlines

In §5.2.1, we showed that thanks to the *FDFI* property of the equilibrium under different deadlines, the equilibrium delays can also be uniquely determined as a function of the arrival time of the first commuter incurring any delay in each origin. Furthermore, if a homogenous distribution of deadlines across origins is assumed, property UE(2) still holds. Hence, the recursive logic of §3.4 can be used to solve the equilibrium freeway problem with different commuter deadlines merge-by-merge as well. This would lead to an exact solution for an homogeneous freeway and approximate for a heterogenous freeway.

5.4 Final remarks

In this chapter, we have shown that the qualitative insights derived in chapter 2 hold for more general instances with an heterogeneous network and different commuters deadlines. Unfortunately, solving these more general cases is notoriously more

difficult.

The analysis should now be extended to incorporate joint route and departure time choice and consider networks with multiple destinations. It is expected that solving the dynamic equilibrium problem exactly will not generally be possible and that heuristics will likely be required. The insights derived in this chapter may help adapt some of these heuristics.

For single destination networks, Akamatsu [2001] recently proposed an efficient algorithm to solve the dynamic equilibrium under route choice and a point queue traffic model. The procedure suggested uses the fact that the solution can be decomposed as a function of the common arrival time to the destination. In §5.3.1, we have shown that, for a two-origin network, it is always possible to formulate the KW-based traffic behavior and the departure-time equilibrium conditions as a function of the arrival time. A similar formulation can be obtained for any single destination network; hence, it seems reasonable to think that Akatmatsu's approach could be naturally extended to consider KW behavior and joint route and departure time choice.

Our results raise more concerns, however, about the possibility to define sound algorithms for multi-destination networks. We showed that the need of preserving FIFO when queues occur in common links leads to equilibrium solutions with discontinuous departure episodes for some origins. At the same time, different types of solution arise depending on the population distributions. Under these circumstances, the equilibrium problem for multi-destination networks may have a combinatorial nature and, in such situations, solutions could only be obtained through *ad-hoc* heuristic procedures. This will not be strange since less restricted problems, which combine both FIFO and multiple destinations, are known to be hard; see Erera et al. [2002]. Huang and Lam [2002] recently presented a simulation-based heuristic to solve joint route

and departure time choice equilibrium in general networks with a point-queue traffic model and a common commuter deadline. Since the algorithm proposed is based on simulating traffic behavior, it could be straightforwardly extended to consider physical queues – by embedding the KW model – and different commuter deadlines. Several issues need to be solved, however. The existence of equilibria and the convergence of the proposed heuristic depend on the continuity and monotonicity of travel times with origin/route inflows [Huang and Lam, 2002; Smith, 1993]. These properties hold with point queues almost always. With physical queues and fixed departure times, however, some studies show that gridlock situations may arise where travel times are not continuous functions of the inflows [Daganzo, 1996], or where the route choice equilibrium can be very sensitive to the input flows and design parameters [Daganzo, 1998]. Additional analysis needs to be pursued to see if similar problems arise when commuters can choose departure time before the heuristics available are extended to consider traffic equilibrium with physical queues.

Chapter 6

Conclusion

This final chapter summarizes the results in this dissertation and proposes ideas for future research.

6.1 Summary

In this dissertation, we have analyzed how traffic congestion develops in urban areas during the morning commute as a function of the spatial distribution of population, and how congestion affects commuter departure time choice.

We argued that the traditional economic models of congestion provide only limited answers to these questions because: (1) they fail to correctly consider the spatial nature of traffic congestion by neglecting the effect of physical queues and merging bottlenecks, and (2) they often overlook dynamic aspects such as the adaptation of commuters departure time choice to (time-varying) queuing conditions.

As a remedy, we proposed a general analysis framework that for the first time incorporates both realistic traffic behavior and departure time adaptation. This framework combines Vickrey's pioneering representation of departure-time choice [Vickrey,

1969] with the spatial model of traffic dynamics in Newell [1993] and the model of merge traffic interactions in Daganzo [1994, 1995a].

The analytical characterization of equilibrium solutions in a stylized two-origin network (chapter 2) enabled us to unveil fundamental insights about the effects of merge interactions and queue spillovers in cases where commuter have different origins. Inefficient congestion levels arise as a result of commuters jockeying for the scarce capacity in common bottleneck queues. Unlike in the traditional common-origin case where congestion accrues to the full population uniformly, a merge bottleneck grants different priorities to different origins population and reduces the undesirable interaction in common queues. Hence, separating the population into different origins often results in reduced total costs. The benefits accrue principally to the less crowded origin. This indicates that efficient ramp metering schemes can be achieved for single-destination freeways when demand is time-elastic. The extended analysis in chapter 5 and appendix A suggests that priority should be given to lower population origins. We also showed that the beneficial effects of merges are partly offset if, as occurs in long freeways, commuters have to share common downstream queues. This implies that reducing the storage capacity of a freeway can reduce total costs for the morning commute if all drivers travel to a common destination. This counterintuitive results suggest that bringing the origins closer to the destination (i.e., decreasing sprawl) not only decreases free-flow travel times, but it also decreases delays.

From a methodological point of view, the insights above allowed us to conclude that previous models of morning commute, which neglect the physical extent of queues and the merging competition, tend to predict wrongly the distribution of cost among origins and substantially overstate total congestion. Fortunately, the improvements proposed in this thesis to incorporate spatial queuing dynamics and merging behavior

do not increase algorithmic complexity.

By exploiting the fundamental properties of the equilibrium solutions for the basic two network model, we were able to extend the analysis to long freeway corridors (chapter 3) and stylized mono-centric cities (chapter 4). In these settings, we characterized the location-based commuting cost as a function of the distribution of population, freeway storage characteristics and ramp priority. Downstream commuters experience reduced cost because they cut into the queues spilling over their access ramps. Far upstream origins, on the other hand, experience the full cost since they always encounter the full queues downstream. This result suggests then that, contrary to current ramp metering practice, it is better not to restrict ramp flows at downstream access points. Additional analysis showed that when a non-congested network of streets exists as an alternative to the freeway, the total level of congestion may indeed increase with respect to a freeway-only situation. The cost increase is observed when users use the street network to access the freeway at less congested origins. Thus, in practical situations, it would be very important to examine how ramp metering or other access control schemes may affect these diversion patterns.

The freeway/city models also confirmed that a larger population sprawl results on more congestion. Nevertheless, the numerical tests indicated that the distribution of population may not be a very important driver of congestion on cities with predominantly mono-centric work trips. Consequently, policies aimed at changing commuter timing behavior (e.g., congestion tolls or access control policies) may be more efficient in general than those aimed at controlling urban growth.

We finally proposed a continuum formulation as an alternative to the network-based (discrete) representation, and obtained approximate closed-form expressions for the commuting costs as a function of distance from the city center and the pop-

ulation distribution (chapter 4, §4.3). These formulae (e.g., (4.17)) characterize in a simple way: (i) the overall system evolution as a function of a few parameters, and (ii) the congestion interaction among populations located at different locations. Furthermore, they provide a consistent structural relationship between cost and population distribution since they incorporate endogenously realistic timing decisions and traffic dynamics. We found that commuting cost at an origin is a weighted function of the cumulative *intervening population* (the population located between the origin and the destination) and the *intervening freeway storage*. They are so simple that they should be very good substitutes for the traditional steady-state representations of congestion costs in urban location models.

To preserve the tractability of most of the analysis in this dissertation, some simplifying assumptions where adopted; namely, the network was homogenous and commuters had the same desired arrival time to the destination. These assumptions were relaxed in chapter 5 and solutions exhibited the same qualitative behavior.

6.2 Future work

This dissertation is not the final word on the analysis of morning commute congestion. Several research areas related to this thesis deserve further investigation.

(a) *Joint route choice/departure time models.* This dissertation examined departure-time choice in single destination networks without route choice. The consideration of simultaneous route and departure time choice (SRD) is a next topic. Some brief remarks on this matter have been outlined in §5.4. The study of algorithms for the SRD problem is a very active research area. Some heuristics algorithms (see, e.g., [Huang and Lam, 2002]) already exists to solve the prob-

lem on general networks with a point-queue traffic model. These algorithms are based on simulating traffic behavior and can be extended to consider physical queues – embedding the KW model – and different commuter deadlines. Still, several issues need to be resolved since the existence of equilibria and the convergence of most heuristics is not guaranteed when spillovers are an issue. Computational efficiency is also important. The results in this dissertation could be used to improve heuristic procedures in special cases. For instance, for many-to-one networks, some route-decomposition strategy may be possible that enables obtaining the equilibrium by solving a sequence of tree-network problems.

(b) *Polycentric scenarios.* Many metropolitan areas are becoming increasingly polycentric nowadays. Although the single destination scenario may still applicable if work is concentrated in a few centers, it is also necessary to analyze cases where the working activity is very disperse. As shown in this dissertation, it is of much value to consider scenarios that can be treated analytically. An approach similar to the one in this dissertation can be applied to the case of *translationally symmetric* cities: cities where the commuter density and the trip distribution length are identical at every location. There is a parallelism between this problem and the monocentric one because commuters only have to be differentiated in one way in both cases: by location in the monocentric city, and by commute trip length in the translationally symmetric city. Timing decisions will depend on trip length instead of location and congestion now will change with time but not with space.

(c) *Spatially-dependent pricing.* Most applied work in congestion pricing only focuses on global efficiency gains. The efficiency gains that can be achieved by

across-the-board tolls, however, are known to have a regressive effect on non-wealthy population groups and this often makes them politically unpopular. More equitable tolling schemes that exempt from the toll a different subset of the vehicular population each day (e.g., on odd or even days, by weekday, etc.) have been devised [Daganzo, 1995b; Daganzo and Garcia, 2000]. These schemes are now possible thanks to electronic tolling technologies, but they have only been studied in simple cases that do not involve the geographical distribution of population. Our equilibrium results strongly suggest that spatial equity should also be considered in the design of tolling schemes, since the desired effects of tolls may differ with location given the distributed impact of queues. Time- and location-dependent tolling strategies that can reduce congestion equitably should be sought.

(d) *Capacity investment vs. sprawl.* The analysis of sprawl in this dissertation has focused exclusively on user cost. A more complete analysis must take in account the investment costs required to provide capacity. Economists have long studied the optimal investment problem under different pricing assumptions but always under unrealistic static scenarios. As noted by Vickrey [1969] early on, capacity is provided mainly to satisfy peak-hour demand; hence the analysis should explicitly consider timing decisions to account for possible shifts in demand and assess the real benefits of the capacity expansion. Not much work is done that combines both timing and spatial effects. The continuous models presented in §4.3 should be a good starting point.

(e) *Urban Location Models Revisited.* The traditional economic equilibrium models of urban location should incorporate traffic models that account for queuing effects. The cost curves developed in this dissertation can be used as a funda-

mental ingredient in the determination of residential location equilibrium. Since the proposed cost expressions are simple (e.g., a linear function of the intervening people) and generally valid, they can add realism without complicating the analysis too much. If the results are simple, further qualitative insights may be gained into the residential location problem.

Bibliography

Akamatsu, T. (2001). An efficient algorithm for dynamic traffic equilibrium assignment with queues, *Trasnportation Science* **35**(4): 389–404.

Akamatsu, T. and Kuwahara, M. (1999). A capacity increasing paradox for a dynamic traffic assignment with departure time choice, *in* A. Ceder (ed.), *Transportation and Traffic Theory (Proceedings of the 14th International Symposium)*, Pergamon Press, Amsterdam, pp. 301–324.

Arnott, R. (2001). The economic theory of urban traffic congestion: a microscopic research agenda, Workshop on Environmental Economics and the Economics of Congestion: Coping with Externalities, Venice International University, Venice Summer Institute, San Servolo, Italy. Found online at http://FMWWWW.bc.edu/EC-V/Arnott.fac.html.

Arnott, R. and MacKinnon, J. G. (1978). Markets and shadow land rents with congestion, *American Economic Review* **68**: 588–600.

Arnott, R., de Palma, A. and Lindsey, R. (1990). Economics of a bottleneck, *Journal of Urban Economics* **27**: 111–130.

Arnott, R., de Palma, A. and Lindsey, R. (1993a). Properties of dynamic traffic equilibrium involving bottlenecks, including a paradox and metering, *Transportation Science* **27**(2): 148–160.

Arnott, R., de Palma, A. and Lindsey, R. (1993b). A structural model of peak-period congestion: A traffic bottleneck with elastic demand, *American Economic Review* **83**: 161–179.

Arnott, R., de Palma, A. and Lindsey, R. (1998). Recent developments in the bottleneck model, *in* K. Button and E. Verhoef (eds), *Road pricing, traffic congestion and the environment: Issues of efficiency and social feasibility*, Edward Elgar, Cheltenham.

Beckmann, M., McGuire, C. and Winsten, C. (1956). *Studies in the Economics of Transportation*, Yale University Press, New Haven, CT.

Bernstein, D., Friesz, T., Tobin, R. and Wie, B. (1993). A variational control formulation of the simultaneous route and departure-time choice equilibrium problem, *in* C. F. Daganzo (ed.), *Transportation and Traffic Theory (Proceedings of the 12th International Symposium)*, Elsevier, New York, pp. 107–124.

Braess, D. (1968). Über ein paradoxen des verkehrsplanung, *Unternehmenforschung* **12**: 258–268.

Branston, D. (1976). Link capacity functions: a review, *Transportation Research* **10**(4): 223–236.

Brockfeld, E., Kuhne, R. D., Skabardonis, A. and Wagner, P. (2003). Towards a benchmarking of microscopic traffic flow models, *Transportation Research Board*. 82nd Annual Meeting, Washington, D.C.

Bureau of Public Roads (1964). *Traffic Assignment Manual*, U.S. Dept. of Commerce, Urban Planning Division, Washington, D.C.

Chu, X. (1992). Endogenous trip scheduling: A comparison of the Vickrey approach and the Henderson approach, *Journal of Urban Economics* **37**: 324–343.

Cox, W. (2000). How urban density intensifies traffic congestion and air pollution, *Arizona Issue Analysis 162*. (accessible at http://www.goldwaterinstitute.org/article.php/95.html).

Dafermos, S. (1980). Traffic equilibrium and variational inequalities, *Transportation Science* **14**: 42–54.

Daganzo, C. F. (1983). Stochastic network equilibrium with multiple vehicle types and asymetric, indefinite link cost jacobians, *Transportation Science* **17**(3): 282–300.

Daganzo, C. F. (1985). The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck, *Transportation Science* **19**(1): 29–37.

Daganzo, C. F. (1994). The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory, *Transportation Research B* **28**: 269–287.

Daganzo, C. F. (1995a). The cell transmission model, part II: Network traffic, *Transportation Research B* **29**: 79–93.

Daganzo, C. F. (1995b). A pareto optimum congestion reduction scheme, *Transportation Research B* **29**(2): 139–154.

Daganzo, C. F. (1995c). Properties of link travel time functions under dynamic loads, *Transportation Research B* **29**: 95–98.

Daganzo, C. F. (1996). The nature of freeway gridlock and how to prevent it, *in* J. Lesort (ed.), *Transportation and Traffic Theory (Proceedings of the 13th International Symposium)*, Pergamon Press, Amsterdam, pp. 629–646.

Daganzo, C. F. (1997). *Fundamentals of Transportation and Traffic Operations*, Elsevier Science, New York, U.S.A.

Daganzo, C. F. (1998). Queue spillovers in transportation networks with a route choice, *Transportation Science* **32**(1): 3–11.

Daganzo, C. F. and Garcia, R. C. (2000). A pareto improving strategy for the time-dependent morning commute problem, *Transportation Science* **34**(3): 303–311.

Daganzo, C. F. and Lin, W.-H. (1994). The spatial evolution of queues during the morning commute in a single corridor, *Research report*, Institute of Transportation Studies, Berkeley, CA.

Daganzo, C. F. and Sheffi, Y. (1977). On stochastic models of traffic assignment, *Transportation Science* **11**: 253–274.

Erera, A. L., Daganzo, C. F. and Lovell, D. J. (2002). The access control problem on capacitated fifo networks with unique o-d paths is hard, *Operations Research* **50**(4): 736–743.

Facchinei, F. and Soares, F. (1995). Testing a new class of algorithms for nonlinear complementarity problems, *in* A. Maugeri and F. Giannessi (eds), *Variational Inequalities and Network Equilibrium Problems*, Plenum Press.

Friesz, T., Luque, J., Tobin, R. and Wie, B. (1989). Dynamic network traffic assignment considered as a continuous time optimal control problem, *Operations Research* **37**: 893–901.

Hau, T. (1998). Congestion pricing and road investment, *in* K. Button and E. Verhoef (eds), *Road pricing, traffic congestion and the environment: Issues of efficiency and social feasibility*, Edward Elgar, Cheltenham.

Henderson, J. (1975). Congestion and optimum city size, *Journal of Urban Economics* **2**: 48–62.

Henderson, J. (1977). *Economic theory and the cities*, Academic Press, New York.

Hendrickson, C. and Kocur, G. (1981). Scheduled delay and departure time in a deterministic models, *Transportation Science* **15**: 62–77.

Huang, H.-J. and Lam, W. H. (2002). Modeling and solving the dynamic equilibrium route and departure time choice problem in networks with queues, *Transportation Research B* **36B**: 253–273.

Janson, B. (1991). Dynamic traffic assignment for urban road networks, *Transportation Research B* **25**: 143–161.

Keeler, T. E. and Small, K. A. (1977). Optimal peak-load pricing, investment and services levels on urban expressways, *Journal of Political Economy* **85**(1): 1–25.

Kuwahara, M. (1990). Equilibrium queuing patterns at a two-tandem bottleneck during the morning peak, *Transportation Science* **24**(3): 217–229.

Kuwahara, M. and Akamatsu, T. (1993). Dynamic equilibrium assignment with queues for one-to-many OD pattern, *in* C. F. Daganzo (ed.), *Transportation and Traffic Theory (Proceedings of the 11th International Symposium)*, Elsevier, New York, N.Y., pp. 185–204.

Kuwahara, M. and Akamatsu, T. (1997). Decomposition of the reactive dynamic assignments with queues for a many-to-many origin-destination pattern, *Transportation Research B* **31**: 1–10.

Kuwahara, M. and Akamatsu, T. (2001). Dynamic user optimal assignment with physical queues for a many-to-many OD pattern, *Trasnportation Research B* **35**: 461–479.

Kuwahara, M. and Newell, G. F. (1987). Queue evolution on freeways leading to a single core city during the morning peak, *in* N. Gartner and N. Wilson (eds), *Transportation and Traffic Theory (Proceedings of the 10th International Symposium)*, Elsevier, New York, pp. 21–40.

Laih, C.-H. (1994). Queueing at a bottleneck with single and multi-step tolls, *Transportation Research A* **28A**(3): 197–208.

Lam, W. and Huang, H. (1995). Dynamic user optimal traffic assignment model for many to one travel demand, *Transportation Research B* **29**: 243–259.

Lighthill, M. and Whitman, G. (1955). On kinematic waves. I flow movement in long rivers, II a theory of traffic flow on long crowded roads, *Proc. Royal Society, A* **229**: 281–345.

Lindsey, R. and Verhoef, E. (2000). Congestion modelling, *in* D. Hensher and K. Button (eds), *Handbook of Transport Modelling*, Elsevier Science, New York.

Lo, H. (1999). A dynamic traffic assignment formulation taht encapsulates the cell-transmission model, *in* A. Ceder (ed.), *Transportation and Traffic Theory (Proceedings of the 14th International Symposium)*, Pergamon Press, Amsterdam, pp. 327–350.

Mahmassani, H. and Herman, R. (1984). Dynamic user equilibrium departures times and route choice on idealized traffic arterials, *Transportation Science* **18**: 362–384.

May, A. M. (1990). *Traffic Flow Fundamentals*, Prentice Hall, Englewood Cliff, NJ.

Mills, E. and de Ferranti, A. (1971). Market choices and optimal city size, *American Economic Review* **61**: 340–345.

Mohring, H. and Harwitz, M. (1962). *Highway benefits: An analytical Approach*, Northwestern University Press, Evanston, IL.

Nagurney, A. and Zhang, D. (1998). Introduction to projected dynamical systems for traffic network equilibrium problems, *in* L. Lundqvist, L.-G. Mattsson and T. Kim (eds), *Network Infrastructure and the Urban Environment; Advances in Spatial Systems Modelling*, Springer, Berlin, pp. 125–156.

Newell, G. F. (1987). The morning commute for non-identical travelers, *Transportation Science* **21**(2): 74–88.

Newell, G. F. (1988). Traffic flow for the morning commute, *Transportation Science* **22**(1): 47–58.

Newell, G. F. (1993). A simplified theory of kinematic waves in highway traffic, I general theory, II queuing a freeway bottlenecks, III multi-destination flows, *Transportation Research B* **27B**(1): 281–313.

Oron, Y., Pines, D. and Sheshinski, E. (1973). Optimum vs. equilibirum patterns and congestion tolls, *Bell Journal of Economics and Management Science* **4**: 619–636.

Papageorgiou, M. (1990). Dynamic modeling, assignment, and route guidance in traffic networks, *Transportation Research B* **24**: 471–495.

Patriksson, M. (1994). *The Traffic Assignment Problem: Models and Methods*, Utrecht, Netherlands.

Pigou, A. (1920). *The Economics of Welfare*, MacMillan and Co., London.

Ran, B. and Boyce, D. (1996). *Modeling Dynamic Transportation Networks: An Intelligent Transportation System Oriented Approach*, Springer, New York.

Ran, B., Boyce, D. and LeBlanc, L. (1993). A new class of instantaneous dynamic user-optimal traffic assignment models, *Operations Research* **41**: 192–202.

Ran, B., Hall, R. and Boyce, D. (1996). A link-based variational inequality model for dynamic departure time/route choice, *Transportation Research B* **30**: 31–46.

Richards, P. (1956). Shockwaves on the highway, *Operations Research* **4**: 42–51.

Ross, S. and Yinger, J. (2000). Timing equilibria in a urban model with congestion, *Journal of Urban Economics* **47**: 390–413.

Small, K. (1982). The scheduling of consumer activities: Work trips, *American Economic Review* **72**: 467–479.

Small, K. (1992). Trip scheduling in urban transportation analysis (in transportation economics), *American Economic Review* **82**(2): 482–486.

Smeed, R. (1967). Some circumstances in which vehicles will reach their destinations earlier by starting later, *Transportation Science* **1**: 308–317.

Smith, M. (1984). The existence of a time-dependent equilibrium distribution of arrivals at a single bottleneck, *Transportation Science* **18**(4): 385–394.

Smith, M. (1993). A new dynamic traffic model and the existence and calculation of dynamic user equilibria on congested capacity-constrained road networks, *Transportation Research B* **27**: 49–63.

Solow, R. (1973). Congestion costs and the use of land for streets, *Bell Journal of Economics and Management Science* **4**: 602–618.

Sullivan, A. (1983a). The general equilibrium effects of congestion externalities, *Journal of Urban Economics* **14**: 80–104.

Sullivan, A. (1983b). Second best policies for congestion externalities, *Journal of Urban Economics* **14**: 105–123.

Tong, C. and Wong, S. (2000). A predictive dynamic traffic assignment in congested capacity-constrained road networks, *Transportation Research B* **34B**: 625–644.

TTI (2002). 2002 urban mobility report, the road information program (trip), *Technical report*, Texas Transportation Institute (TTI).

UNDP (1997). International survey of mayors (for the conference on governance for sustainable growth and equity), *Technical report*, United Nations Development Program, UN headquarters. (info available at http://magnet.undp.org/Docs/urban/Maysur.htm).

Vickrey, W. (1969). Congestion theory and transport investment, *American Economic Review* **59**: 251–260.

Walters, A. (1961). The theory and measurement of private and social cost of highway congestion, *Econometrica* **29**(4): 676697.

Wardrop, J. (1952). Some theoretical aspects of road traffic research, *Proceedings of the Institute of Civil Engineers* **1**(II): 325–378.

Wheaton, W. (1998). Land use and density in cities with congestion, *Journal of Urban Economics* **43**: 258–272.

Wie, B., Friesz, T. and Bernstein, D. (1995). A discrete time, nested cost operator approach to the dynamic network user equilibrium problem, *Transportation Science* **29**(1): 79–92.

Wie, B., Friesz, T. and Tobin, R. (1990). Dynamic user optimal traffic assignment on congested multidestination networks, *Transportation Research B* **24B**: 431–442.

Wie, B.-W. and Tobin, R. (1998). Dynamic congestion pricing models for general traffic networks, *Transportation Research B* **32**: 313327.

Windover, J. (1998). *Empirical Studies of the Dynamic Features of Freeway Traffic*, PhD thesis, Department of Civil and Environmental Engineering, University of California, Berkeley, CA.

Yinger, J. (1993). Bumper to bumper: A new appraoch to congestion in an urban model, *Journal of Urban Economics* **34**: 249–274.

Appendix A

Two-Origin Network: Metering and Capacity Expansion

In this appendix, we analyze the optimal ramp metering and capacity expansion strategies for the two-origin heterogenous network of section 5.1. Without loss of generality, we adopt: q_{MD} (the downstream link capacity) as the unit of capacity, hence the capacities at the upstream links are defined as the ratios $\zeta^{(A)} = q_{AM}/q_{MD}$ and $\zeta^{(B)} = q_{BM}/q_{MD}$. Cost is expressed in units of $\frac{1}{q_{MD}}\frac{eL}{e+L}$ (the single bottleneck equilibrium cost with population 1).

No spillovers

Consider first the solution with no spillovers of Figure 5.2. Then the following properties hold:

Property A.1. (*Optimal Ramp Metering*) If we assume fixed link capacities and that variable priority ratios $\tilde{\alpha}^{(A)}$ and $\tilde{\alpha}^{(B)}$ as a function of some metering control, then the optimal metering is always $(\tilde{\alpha}^{(A)}, \tilde{\alpha}^{(B)}) = (\zeta^{(A)}, 1-\zeta^{(A)})$ or $(\tilde{\alpha}^{(A)}, \tilde{\alpha}^{(B)}) =$

$(1-\zeta^{(B)}, \zeta^{(B)})$.

Proof. First notice that we should always consider priority ratios such that $\tilde{\alpha}^{(A)} \leq \zeta^{(A)}$ and $\tilde{\alpha}^{(B)} \leq \zeta^{(B)}$ since, otherwise, the full capacity downstream is never used and the total cost can always be reduced by using the full capacity. Hence, for any population split $\{N^{(A)}, N^{(B)}\}$, we should look for the optimal $\tilde{\alpha}^{(A)} \in [1-\zeta^{(B)}, \zeta^{(A)}]$ (or equivalently, $\tilde{\alpha}^{(B)} \in [1-\zeta^{(A)}, \zeta^{(B)}]$).

Each origin equilibrium cost varies continuously with $\tilde{\alpha}^{(A)}$. When $\frac{N^{(A)}}{\tilde{\alpha}^{(A)}} > \frac{N^{(B)}}{\tilde{\alpha}^{(B)}}$, (i.e., $\frac{N^{(A)}}{N^{(A)}+N^{(B)}} > \tilde{\alpha}^{(A)}$), they are given by

$$\tilde{C}^{(A)} = \left(\frac{N^{(A)}}{\zeta^{(A)}} + \frac{N^{(B)}}{\tilde{\alpha}^{(B)}} (1 - \frac{\tilde{\alpha}^{(A)}}{\zeta^{(A)}}) \right) \frac{eL}{e+L} \text{ and } \tilde{C}^{(B)} = \frac{N^{(B)}}{\tilde{\alpha}^{(B)}}. \tag{A.1}$$

Then,

$$\frac{\mathrm{d}TC}{\mathrm{d}\tilde{\alpha}^{(A)}} = N^{(A)} \frac{\mathrm{d}\tilde{C}^{(A)}}{\mathrm{d}\tilde{\alpha}^{(A)}} + N^{(B)} \frac{\mathrm{d}\tilde{C}^{(B)}}{\mathrm{d}\tilde{\alpha}^{(A)}} = \left((\frac{N^{(B)}}{\tilde{\alpha}^{(B)}})^2 \left\{ \frac{N^{(A)}}{N^{(B)}} \left(1 - \frac{\tilde{\alpha}^{(B)}}{\zeta^{(A)}} + \frac{\tilde{\alpha}^{(A)}}{\zeta^{(A)}} \right) \right\} + (\frac{N^{(B)}}{\tilde{\alpha}^{(B)}})^2 \right) \tag{A.2}$$

and, since $\tilde{\alpha}^{(A)} + \tilde{\alpha}^{(B)} = 1$, we have

$$\frac{\mathrm{d}TC}{\mathrm{d}\tilde{\alpha}^{(A)}} = (\frac{N^{(B)}}{\tilde{\alpha}^{(B)}})^2 \left\{ \frac{N^{(A)}}{N^{(B)}} \left(1 - \frac{1}{\zeta^{(A)}} \right) + 1 \right\}. \tag{A.3}$$

Equivalently, when $\frac{N^{(A)}}{\tilde{\alpha}^{(A)}} < \frac{N^{(B)}}{\tilde{\alpha}^{(B)}}$, we have

$$\frac{\mathrm{d}TC}{\mathrm{d}\tilde{\alpha}^{(A)}} = (\frac{N^{(A)}}{\tilde{\alpha}^{(A)}})^2 \left\{ \frac{N^{(B)}}{N^{(A)}} \left(1 - \frac{1}{\zeta^{(B)}} \right) + 1 \right\}. \tag{A.4}$$

From (A.3) and (A.4), then we have

$$\frac{\mathrm{d}TC}{\mathrm{d}\tilde{\alpha}^{(A)}} = \begin{cases} > 0 & \text{if } 0 \leq \frac{N^{(A)}}{N^{(A)}+N^{(B)}} \leq 1-\zeta^{(B)} \\ < 0 & \text{if } 1-\zeta^{(B)} < \frac{N^{(A)}}{N^{(A)}+N^{(B)}} \leq \tilde{\alpha}^{(A)} \\ > 0 & \text{if } \tilde{\alpha}^{(A)} < \frac{N^{(A)}}{N^{(A)}+N^{(B)}} \leq \zeta^{(A)} \\ < 0 & \text{if } \zeta^{(A)} < \frac{N^{(A)}}{N^{(A)}+N^{(B)}} \leq 1 \end{cases}. \tag{A.5}$$

The optimal $\tilde{\alpha}^{(A)}$ is $1-\zeta^{(B)}$ when $0 \leq \frac{N^{(A)}}{N^{(A)}+N^{(B)}} \leq 1-\zeta^{(B)}$ and $\zeta^{(A)}$ when $\zeta^{(A)} < \frac{N^{(A)}}{N^{(A)}+N^{(B)}} \leq 1$. When $1-\zeta^{(B)} \leq \frac{N^{(A)}}{N^{(A)}+N^{(B)}} \leq \zeta^{(A)}$ both $\tilde{\alpha}^{(A)} = \zeta^{(A)}$ and $\tilde{\alpha}^{(A)} = 1-\zeta^{(B)}$ are optimal since they yield the same total minimum optimal cost $TC^* = \frac{{N^{(A)}}^2}{\zeta^{(A)}} + \frac{{N^{(B)}}^2}{\zeta^{(B)}} + \frac{N^{(A)}N^{(B)}}{\zeta^{(A)}\zeta^{(B)}}(\zeta^{(A)}+\zeta^{(B)}-1)$.

An alternative graphical argument maybe more intuitive. Consider the arrival-time domains $\Pi_q = \{t : d_D(t) = q_M D = 1\}$ and $\Pi = \{t : d_D(t) > 0\}$ in Figure 5.2. Obviously, $\Pi_q \subset \Pi$. Let $d\tilde{C}^{(B)}$ be the infinitesimal change in $\tilde{C}^{(B)}$ due to a change in priority $\zeta^{(A)}$. Then it is straightforward from the figure that

$$d|\Pi_q| = dt_s^{(B)} + dt_f^{(B)} = \frac{1}{e}d\tilde{C}^{(B)} + \frac{1}{L}d\tilde{C}^{(B)} = \frac{e+L}{eL}d\tilde{C}^{(B)} \tag{A.6}$$

$$d|\Pi| = \left(1-\frac{1}{\zeta^{(A)}}\right)d|\Pi_q| = \left(1-\frac{1}{\zeta^{(A)}}\right)\frac{e+L}{eL}d\tilde{C}^{(B)} \tag{A.7}$$

$$d\tilde{C}^{(A)} = \frac{eL}{e+L}d|\Pi| = \left(1-\frac{1}{\zeta^{(A)}}\right)d\tilde{C}^{(B)} \tag{A.8}$$

Then,

$$\frac{\mathrm{d}TC}{\mathrm{d}\tilde{\alpha}^{(A)}} = N^{(A)}\frac{\mathrm{d}\tilde{C}^{(A)}}{\mathrm{d}\tilde{\alpha}^{(A)}} + N^{(B)}\frac{\mathrm{d}\tilde{C}^{(B)}}{\mathrm{d}\tilde{\alpha}^{(A)}} = \frac{\mathrm{d}\tilde{C}^{(B)}}{\mathrm{d}\tilde{\alpha}^{(A)}}\left\{N^{(A)}\left(1-\frac{1}{\zeta^{(A)}}\right)+N^{(B)}\right\} \tag{A.9}$$

Hence, we recover the same criteria as before. ■

The results in property A.1 are summarized in Figure A.1. The continuous bold lines represent the optimal metering as a function of the ratio $\frac{N^{(A)}}{N^{(A)}+N^{(B)}}$. Note that property A.1 confirms the idea that under departure time choice the optimal ramp metering must give to one of the approaches as much priority as possible as long as the downstream link is not starved. Since this is not generally feasible and one can only alter the natural merge priority moderately, what the analysis suggest is that one

should give at least some additional priority to the approach with the lower *priority-to-population* ratio (i.e., ramps in the freeway). Furthermore, one should only meter during the rush interval where queues exist in both approaches.

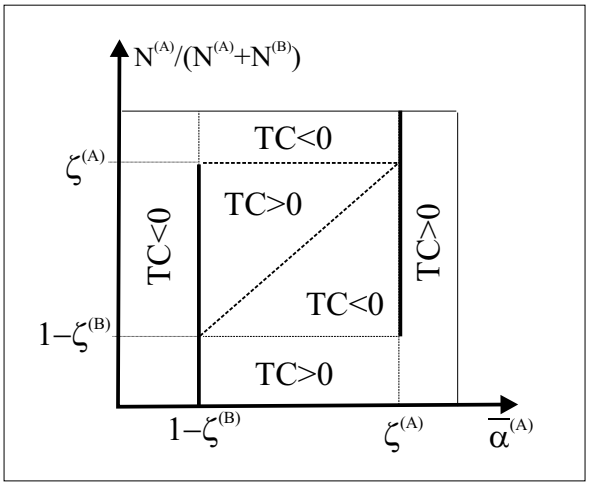

Figure A.1. Total cost change with metering.

Property A.2. (*Capacity investment*). Assume that the priority ratios are proportional to each approach capacity, i.e., $\tilde{\alpha}^{(A)}/\tilde{\alpha}^{(B)} = \zeta^{(A)}/\zeta^{(B)}$ and A is the higher *population-to-priority* approach (i.e., $\frac{N^{(A)}}{N^{(B)}} > \frac{\tilde{\alpha}^{(A)}}{\tilde{\alpha}^{(B)}}$), then,

(a) An increase in the capacity downstream capacity q_{MD} always reduces total cost.

(b) System cost decreases with an increase in the capacity of approach B as long as $1 - \zeta^{(A)} < \frac{N^{(B)}}{N^{(A)}+N^{(B)}} < \tilde{\alpha}^{(B)}$.

(c) System cost always decreases with an increase in the capacity of A.

Proof.

a This is immediate since both $\tilde{C}^{(A)}$ and $\tilde{C}^{(B)}$ are inversely proportional to q_{MD}. (Note that $\tilde{C}^{(A)}$ and $\tilde{C}^{(B)}$ in (A.1) are expressed in units of $\frac{1}{q_{MD}}\frac{eL}{e+L}$.)

b This is also immediate since when B is the lower *population-to-priority* origin, the total cost is only affected through $\tilde{\alpha}^{(B)}$. Increasing $\zeta^{(B)}$ increases $\tilde{\alpha}^{(B)}$, and then by virtue of property A.1, the total cost decrease always when $1-\zeta^{(A)} < \frac{N^{(B)}}{N^{(A)}+N^{(B)}} < \tilde{\alpha}^{(B)}$.

c First consider that $\frac{\mathrm{d}\tilde{\alpha}^{(A)}}{\mathrm{d}\zeta^{(A)}} = \frac{\zeta^{(B)}}{(\zeta^{(A)}+\zeta^{(B)})^2}$ and $\frac{\mathrm{d}\tilde{\alpha}^{(B)}}{\mathrm{d}\zeta^{(A)}} = -\frac{\zeta^{(B)}}{(\zeta^{(A)}+\zeta^{(B)})^2}$. Then, taking derivatives with respect to $\zeta^{(A)}$ in (A.1), we have

$$\frac{\mathrm{d}\tilde{C}^{(A)}}{\mathrm{d}\zeta^{(A)}} = -\frac{N^{(A)}}{{\zeta^{(A)}}^2} + \frac{N^{(B)}}{\tilde{\alpha}^{(B)}};\ \frac{\mathrm{d}\tilde{C}^{(B)}}{\mathrm{d}\zeta^{(A)}} = \frac{N^{(B)}}{\zeta^{(B)}}. \tag{A.10}$$

Hence,

$$\frac{\mathrm{d}TC}{\mathrm{d}\zeta^{(A)}} = -\frac{{N^{(A)}}^2}{{\zeta^{(A)}}^2} + \frac{N^{(B)}}{\tilde{\alpha}^{(B)}}\left(N^{(A)}+N^{(B)}\right). \tag{A.11}$$

From (A.11), $\frac{\mathrm{d}TC}{\mathrm{d}\zeta^{(A)}} < 0 \iff \frac{N^{(A)}}{N^{(A)}+N^{(B)}} > \frac{\zeta^{(A)}}{\zeta^{(A)}+\zeta^{(B)}}$. But $\frac{N^{(A)}}{N^{(B)}} > \frac{\tilde{\alpha}^{(A)}}{\tilde{\alpha}^{(B)}} = \frac{\zeta^{(A)}}{\zeta^{(B)}}$ implies $\frac{N^{(A)}}{N^{(A)}+N^{(B)}} > \frac{\zeta^{(A)}}{\zeta^{(A)}+\zeta^{(B)}}$, hence $\frac{\mathrm{d}TC}{\mathrm{d}\zeta^{(A)}} < 0$. ■

Property 2 suggest that a capacity increasing paradox (i.e., total cost increasing when more capacity is provided) is not likely, since one would normally increase the capacity of the more congested approach.

Spillovers

The analysis of optimal metering and capacity investment complicates when spillovers exists since equilibrium patterns can be very diverse depending on the existing restrictions at D. Here we develop some basic guidelines considering the case of a fixed capacity bottleneck at D with $\zeta_D = q_D/q_{MD} < 1$; see Figure 5.3.

Property A.3. Assuming A is the origin with higher *population-to-priority* , then:

(a) If $\zeta^{(A)} > \zeta_D$, then it is always optimal to meter the combined flows at the merge, so that no queues develop at D (i.e., $q_M(t) \leq q_D$), and give as much priority as

possible to approach B.

(b) If $\zeta^{(A)} < \zeta_D$, then metering the combined flows at M (i.e., $q_M(t) \leq q_D$) and giving as much priority to B as possible is still beneficial as long as $\tilde{\alpha}^{(A)} < \frac{N^{(A)}}{N^{(B)}} < \frac{\zeta^{(A)}}{1-\zeta^{(A)}} \frac{e+L}{L}$.

Proof.

a This immediate from Figure 5.3b,d since reducing spillovers and giving priority to B reduces $\tilde{C}^{(B)}$ but does not affect $\tilde{C}^{(A)}$. Actually, we can interpret this problem as having a homogeneous network problem with link capacities q_D and a 'generalized' spillover curve. Then, as in the homogeneous case solution, it is always optimal to avoid spillovers.

b We use the same graphical logic of property A.1. Consider the arrival-time domains $\Pi_q = \{t : d_D(t) = q_D\}$ and $\Pi = \{t : d_D(t) > 0\}$ in Figure 5.3a,c and $d\tilde{C}^{(B)}$ be the infinitesimal change in $\tilde{C}^{(B)}$ due to any change due to metering. In this case,

$$d|\Pi_q| = dt_s^{(B)} = \frac{1}{e} d\tilde{C}^{(B)} \tag{A.12}$$

$$d|\Pi| = \left(1 - \frac{1}{\zeta^{(A)}}\right) d|\Pi_q| = \left(1 - \frac{1}{\zeta^{(A)}}\right) \frac{1}{e} d\tilde{C}^{(B)} \tag{A.13}$$

$$d\tilde{C}^{(A)} = \frac{eL}{e+L} d|\Pi| = \left(1 - \frac{1}{\zeta^{(A)}}\right) \frac{L}{e+L} d\tilde{C}^{(B)} \tag{A.14}$$

Then,

$$dTC = d\tilde{C}^{(B)} \left\{ N^{(A)} \left(1 - \frac{1}{\zeta^{(A)}}\right) \frac{L}{e+L} + N^{(B)} \right\} \tag{A.15}$$

Since reducing spillovers or giving priority to B makes $d\tilde{C}^{(B)} < 0$, then $dTC < 0 \iff \tilde{\alpha}^{(A)} < \frac{N^{(A)}}{N^{(B)}} < \frac{\zeta^{(A)}}{1-\zeta^{(A)}} \frac{e+L}{L}$. ■

Note that the situation in 1 is typical of a bottleneck occurring downstream of a ramp merge (i.e., the capacity of the upstream freeway – approach A – is always larger than the bottleneck – D). Our results suggest that avoiding queues in the downstream freeway, which is one of the main objectives of freeway ramp metering should be beneficial. However, this should be done by restricting upstream freeway flows rather than ramp flows. Since metering freeway is not possible, one would like at least not to meter the ramp.

Appendix B

Nomenclature

What follows is a *partial* list of symbols used in the dissertation. Symbols are categorized by the model they refer to: first, the two-origin network (chapters 2, 5); then, the freeway-city model (chapters 3, 4). A lower-case variable following an upper-case variable represents the time-derivative of the latter.

Chapters 2,5: Two origin-network

$r = A, B$	origins
$A^{(r)}$, $a^{(r)}$	cumulative departures for origin r, inflow
$\tilde{C}^{(r)}$	origin-r equilibrium cost
$D_D^{(r)}$, $d_D^{(r)}$	origin-r cumulative departures from point D, outflows (or arrivals to the destination)
D_D, d_D	aggregated cumulative departure from D (outflow)
D_M, d_M	aggregated cumulative departures from M (outflow)
D_M^D, d_M^D	maximum cumulative allowed departures from M as determined by downstream conditions ("spillover curve")
d_M^{max}	maximum departure rate from M

d_{rM}^{max}	maximum departure rate from M coming from approach rM
k_j	freeway jam density (or storage per unit length)
ℓ	length of link MD
$N^{(r)}$	origin r population
$\tilde{N}^{(r)}$	population threshold for different equilibrium solutions
N_U	total population crossing the merge when it is non-congested
$\eta^{(r)}$	origin r population ratio
q_{max}	freeway capacity and jam density (homogeneous case)
q_D	capacity at point D (possibly time-dependent)
q_M	capacity at point M (time-dependent)
q_{MD}, q_{rM}	link-MD, -rM capacities (heterogeneous case)
$T^{(r)}$	origin-r arrival/departure (A/D) schedule curve
$t_s^{(r)}$	time when the first commuter in origin r may experience delay
$t_f^{(r)}$	time when the last commuter in origin r may experience delay
v_f	freeway free-flow speed
$W^{(r)}$, $w^{(r)}$	cumulative desired arrivals at the destination (*deadline curve*)
w_t	deadline of commuter arriving at time t
$\tilde{\alpha}^{(r)}$	congested ramp priority ratio for approach/origin r
$\alpha_D^{(r)}$	origin-r flow proportion departing from D
$\alpha_M^{(r)}$	origin-r flow proportion departing from M
$\tau^{(r)}$	origin-r equilibrium delay (or trip-time)
τ_{MD}	delay in link MD
ν_{MD}	maximum possible delay (M-delay) in link MD
$\Pi^{(r)}$	arrival time interval for origin r

Chapters 3,4: Linear freeway/Monocentric city model

$r = 1 \dots R$	set of origin
$A^{(r)}$, $a^{(r)}$	cumulative desired departures, inflow, from origin r
A_{ij}, a_{ij}	cumulative arrivals (inflow) to link (i, j)
A_{ij}^D	cumulative allowed arrivals into link (i, j) by downstream spillovers (i.e., "spillover curve" at node i from node j at link (i, j)
A_{ij}^U	cumulative allowed arrivals to link (i,j) in the absence of spillovers (i.e.,"point bottleneck" flows at node i)
a_{ij}^{max}	maximum arrival rate at link (i, j)
$\tilde{C}^{(r)}$	origin-r equilibrium cost
$\tilde{C}(x)$	location-x equilibrium cost (continuous case)
D_{ij}, d_{ij}	cumulative departures (outflow) from link (i, j)
D_{rr}, d_{rr}	cumulative departures (outflow) from ramp r
d_{ij}^{max}	maximum departure rate from link (i, j)
ℓ_{ij}	length of link (i, j)
$\eta^{(r)}$	population at origin r
$\eta(x)$	population density at location x (continuous model)
$\eta_\alpha(x)$	modified population density at location x (continuous model)
N_i	aggregated population originating upstream of merge i
$N(x)$	cumulative population between O and x or *intervening* population (continuous model)
q_{max}, k_j	freeway capacity and jam density
$Y^{(r)}$, $y^{(r)}$	origin-r cumulative arrivals, outflow, to the destination O
Y_i	cumulative departure from link (i, j) (i.e., passing node i) arriving at O by time t.
$\alpha^{(r)}$	congested ramp priority ratio

$\alpha(x)$	ramp priority per unit length at location x (continuous model)
α_{ij}	actual flow share for incoming link (i,j) at merge (i,j)-(i',j)
$\tilde{\alpha}_{ij}$	congested capacity share for incoming link (i,j) at merge (i,j)-(i',j)
$\tau^{(r)}$	origin-r equilibrium delay (or trip-time)
τ_i	freeway delay (or trip-time) from node i to O
τ_i^M	maximum equilibrium delay experienced by users upstream of node i
ΔT_{ij}	additional travel time to travel from origin i to O through origin j
ΔT_{i0}	additional travel time to travel from origin i to O on the street network
$\Pi^{(r)}$	arrival time interval for origin r