Copyright
by
Stephen David Boyles
2009

The Dissertation Committee for Stephen David Boyles
certifies that this is the approved version of the following dissertation:

Operational, Supply-Side Uncertainty in Transportation Networks: Causes, Effects, and Mitigation Strategies

Committee:

S. Travis Waller, Supervisor

Anant Balakrishnan

Chandra Bhat

Randy Machemehl

Zhanmin Zhang

Athanasios Ziliaskopoulos

Operational, Supply-Side Uncertainty in Transportation Networks: Causes, Effects, and Mitigation Strategies

by

Stephen David Boyles, B.S., B.S.C.E., M.S.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2009

In loving memory of May Ying Chin Lee

Acknowledgments

My graduate studies have benefitted greatly from the assistance and guidance of a number of people, both in the University and elsewhere. First and foremost, I must express appreciation to Travis Waller, who supervised my M.S. and Ph.D., introduced me to network analysis, and convinced me to pursue doctoral studies. Thanks are also due to the members of my committee — Anant Balakrishnan, Chandra Bhat, Randy Machemehl, Zhanmin Zhang, and Thanasis Ziliaskopoulos — not only for their advice on this dissertation, but for teaching courses which have been immensely valuable to me and others. Mark Hallenbeck has also served as a mentor to me since taking his transportation planning course at the University of Washington. He has always been willing to speak with me, meet with me, and share data with me, and I greatly appreciate his perspective on matters.

I am also grateful for the cameraderie and *esprit de corps* in the transportation engineering department, which has been a source of friendship and inspiration, in addition to giving me an outlet for nerdy transportation jokes. I would like to call particular attention to the members of the TeQson Lab, to my fellow Ph.D. students who have seen many others in our program come and go, and to the administrative assistants who make sure nobody misses a deadline — Libbie Toler, Chandra Lownes, Lisa Macias, and Vicki Simpson

have been particularly helpful to me in my time here.

My family has been incredibly supportive of me throughout my life. My parents have been unwavering in their love and encouragement for as long as I can remember, and no words can express my gratitude towards them. I have many aunts, uncles, and cousins who have always included me in family events and made me feel welcome when I moved halfway across the country for graduate school. I also thank my grandparents, Franklin and Leslie Boyles, and my late grandmother, May Ying Chin Lee, to whom this dissertation is dedicated. Your constant love, and your example of hard work and discipline, have inspired me countless times, and I look to you.

Last, but not least, I have been blessed to be a member of University United Methodist Church for my time in Austin. This is especially true of the choir and the campus ministry, and I thank Marc Erck and Bill Frisbie for their respective leadership in these groups. Regarding frequent inquiries by the latter as to whether synchronized signals are worth the cost, I must unfortunately report that a satisfactory answer is beyond the scope of this humble dissertation.

Operational, Supply-Side Uncertainty in Transportation Networks: Causes, Effects, and Mitigation Strategies

Publication No. ____________

Stephen David Boyles, Ph.D.
The University of Texas at Austin, 2009

Supervisor: S. Travis Waller

This dissertation is concerned with travel time uncertainty in transportation networks due to ephemeral phenomena such as incidents or poor weather. Such events play a major role in nonrecurring congestion, which is estimated to comprise between one-third and one-half of all delay on freeways. Although past research has considered many individual aspects of this problem, this dissertation is unique in bringing a comprehensive approach, beginning with study of its causes, moving to discussion of its effects on traveler behavior, and then demonstrating how these models can be applied to mitigate the effects of this uncertainty.

In particular, two distinctive effects of uncertainty are incorporated into all aspects of these models: *nonlinear traveler behavior*, encompassing risk aversion, schedule delay, on-time arrival, and other user objectives that explicitly recognize travel time uncertainty; and *information and adaptive routing*, where travelers can adjust their routes through the network as they acquire information on its condition.

In order to accurately represent uncertain events in a mathematical model, some quantitative description of these events and their impacts must be available. On freeways, a large amount of travel data is collected through intelligent transportation systems (ITS), although coverage is far from universal, and very little data is collected on arterial streets. This dissertation develops a statistical procedure for estimating probability distributions on speed, capacity, and other operational metrics by applying regression to locations where such data is available. On arterials, queueing theory is used to develop novel expressions for expected delay conditional on the signal indication.

The effects of this uncertainty are considered next, both at the individual (route choice) and collective (equilibrium) levels. For individuals, the optimal strategy is no longer a path, but an adaptive policy which allows for flexible re-routing as information is acquired. Dynamic programming provides an efficient solution to this problem. Issues related to cycling in optimal policies are examined in some depth. While primarily a technical concern, the presence of cycling can be discomforting and needs to be addressed.

When considering collective behavior, the simultaneous choices of many self-optimizing users (who need not share the same behavioral objective) can be expressed as the solution to a variational inequality problem, leading to existence and uniqueness results under certain regularity conditions. An improved policy loading algorithm is also provided for the case of linear traveler behavior.

Finally, three network improvement strategies are considered: locating information-providing devices; adaptive congestion pricing; and network design. Each of these demonstrates how the routing and equilibrium models can be applied, using small networks as testbed locations. In particular, the information provision and adaptive congestion pricing strategies are extremely difficult to represent without an adaptive equilibrium model such as the one provided in this dissertation.

Table of Contents

Acknowledgments **v**

Abstract **vii**

List of Tables **xiii**

List of Figures **xv**

Chapter 1. Introduction **1**

1.1 Background 1
1.2 Motivation 4
1.3 Problem Statements 7
1.4 Notation 13
1.5 Contributions 17
1.6 Organization 19

Chapter 2. Causes: Determining Travel Time Distributions **23**

2.1 Introduction 23
2.2 Literature Review 25
2.3 Method 27
  2.3.1 Supply-Side Uncertainty on Freeways 28
  2.3.2 Supply-Side Uncertainty on Arterials 32
  2.3.3 Demand-Side Uncertainty 38
2.4 Demonstration 43
  2.4.1 Dallas 43
  2.4.2 Seattle 48
2.5 Conclusion 54

Chapter 3. Effects: Routing and Equilibrium 56

3.1 Introduction 56

3.2 Routing 57

3.2.1 Literature Review 58

3.2.2 Disutility Functions 61

3.2.3 Solution Algorithm 69

3.2.4 Determining Node and Arc Usages 73

3.2.5 Cycles and *Contretemps* 76

3.2.6 A Small Demonstration 84

3.3 Equilibrium 92

3.3.1 Literature Review 95

3.3.2 Extensions for the Linear Case 98

3.3.3 General Properties 104

3.3.4 Solution Algorithms 106

3.3.5 A Small Demonstration 109

3.4 Conclusion 114

Chapter 4. Analysis of Routing and Equilibrium Algorithms 117

4.1 Introduction 117

4.2 Routing Tests 118

4.3 Equilibrium Tests 129

Chapter 5. Improvement Strategies 136

5.1 Introduction 136

5.2 Information Provision 137

5.2.1 Problem Definitions 140

5.2.2 Network Contraction 145

5.2.3 Solution Methods 152

5.2.4 Demonstration 158

5.2.5 Conclusion 165

5.3 Congestion Pricing 165

5.3.1 Literature Review 167

5.3.2 Problem Statement 172

5.3.2.1 Information Scenarios 173
5.3.2.2 User Behavior 175
5.3.2.3 Network Manager Behavior 178
5.3.3 Solution Methods 178
5.3.3.1 Adaptive Routing/Unresponsive Tolls 179
5.3.3.2 Adaptive Routing/Responsive Tolls 182
5.3.3.3 No Information/Unresponsive Tolls 183
5.3.4 Heterogeneous Users 183
5.3.5 Demonstration 189
5.3.6 Conclusion 195
5.4 Network Design 201

Chapter 6. Conclusion **210**
6.1 Summary and Implications 210
6.2 Future Work 214

References **216**

Vita **235**

List of Tables

1.1 Taxonomy describing sources of uncertainty in transportation systems, with examples. 6
1.2 Joint states for demonstration of information concepts. 16
1.3 Conditional probabilities for demonstration of information concepts. 16

2.1 Example ranking of distributions for one detector. 45
2.2 Regression results for NIGW scenario 47
2.3 Regression results for PW scenario. 47
2.4 Regression results for IP scenario. 47
2.5 Seattle data highway segments and network arcs. 49
2.6 Regression results for Seattle data. 53
2.7 Seattle data highway segments and network arcs. 55

3.1 Data for NL-OSP demonstration on Seattle network 87
3.2 Final node labels for NL-OSP in Seattle 88
3.3 Final policy for NL-OSP in Seattle 89
3.4 Node arrivals for NL-OSP in Seattle 90
3.5 Final arc usages for NL-OSP in Seattle 93
3.6 Policies for NL-UER demonstration 111
3.7 Demonstration of NL-UER 115

4.1 Characteristics of the test networks 118
4.2 Computation time for NL-OSP 119
4.3 Results for linear disutility function 122
4.4 Results for deviance disutility function 123
4.5 Results for arriving-on-time disutility function 124

5.1 Node states and optimal policy for the example network with $R = \{B\}$ 142

5.2 Node states and optimal policy for the example network with $R = \{A\}$ 143
5.3 Overview of problems IIP, UIP, and CIP 144
5.4 Individual information provision on test networks 161
5.5 Uncongested information provision on test networks 163
5.6 Congested information provision on test networks 164
5.7 Four paths used in tricriterion demonstration 185
5.8 Comparison of average trip characteristics for all information scenarios 192
5.9 Comparison of system states for all information scenarios . . . 193
5.10 Results from network design on Sioux Falls 207

List of Figures

1.1 Demonstration of information concepts. 14
1.2 Schematic of dissertation components. 22

2.1 Exploding an intersection node to represent turning movements. 33
2.2 Queuing diagram for undersaturated lane groups. 35
2.3 Queuing diagram for oversaturated lane groups. 37
2.4 Comparison of intrinsic and extrinsic sensitivity. 40
2.5 Conditional and unconditional speed distributions. 48
2.6 Seattle network used to demonstrate dissertation algorithms. . 50

3.1 Linear disutility function vs. deviance. 63
3.2 Linear disutility function vs. upside deviance. 64
3.3 Various quadratic disutility functions. 66
3.4 Various Box-Cox disutility functions. 67
3.5 Illustration of a *contretemps* 77
3.6 Demonstration network for NL-UER. 112

4.1 Expected policy cost with varying information levels, Sioux Falls. 126
4.2 Computation time with information level, Sioux Falls. 127
4.3 Computation time with information level, Anaheim. 128
4.4 Computation time with information level, Barcelona. 128
4.5 Computation time with information level, Chicago Regional. . 129
4.6 Convergence of NL-UER for linear-disutility travelers. 130
4.7 Convergence of NL-UER for quadratic-disutility travelers. . 131
4.8 Convergence of NL-UER for mixed travelers. 132
4.9 Comparing user class scenarios for NL-UER 133
4.10 The impact of information provision on NL-UER 135

5.1 Example network to demonstrate notation and concepts. . . . 141

5.2 Example contracted network with $k = 2$. 147
5.3 Sioux Falls network and a contracted graph for two information nodes. 148
5.4 Potential conflict with expanding policies on the contracted graph, and resolution procedure. 151
5.5 Reduction from the 0-1 knapsack problem. 153
5.6 Network demonstrating how local search and the greedy heuristics can fail. 157
5.7 Four efficient paths and their convex hull 186
5.8 Regions of $VOTT$-$VOTR$ space corresponding to each efficient path (the “dual graph”) 187
5.9 Sioux Falls test network; dashed lines indicate degradable (freeway) arcs. 191
5.10 Untolled volume-to-capacity ratios for no-information case . . 196
5.11 Untolled volume-to-capacity ratios for full-information case . . 197
5.12 Change in v/c ratios under tolling (NI/UT); average arc tolls shown in dollars 198
5.13 Change in v/c ratios under tolling (AR/RT); average arc tolls shown in dollars 199
5.14 Change in v/c ratios under tolling (AR/UT); average arc tolls shown in dollars 200

Chapter 1

Introduction

1.1 Background

Transportation systems connect virtually every aspect of modern life. Any consumable item must travel from where it is produced to the end user, often depending on complicated supply chains involving the transport of other goods at a national or global scale. Any movement of people, whether commuting to work, traveling to the store, or visiting out-of-town relatives on vacation, by defintion involves one or more transportation modes and routes. It is no exaggeration to say that the economy of modern nations depends vitally on the ability of their transportation systems to move people and goods efficiently and safely.

Providing this efficient and safe movement is no easy task, and the litany of present and future challenges is well-known. Congested freeways and airports waste countless hours of travelers' time, and cost billions of dollars annually in lost time and wasted fuel. The need to consider the environmental impact of transportation systems is more critical than ever before. Aging infrastructure will require novel funding mechanisms to ensure the continued safety and reliability of structures and pavements themselves. New needs have also been discovered, such as ensuring equitable distribution of the costs and benefits of transportation improvements. A multitude of new electronic and communication technologies, under the umbrella of intelligent transportation systems (ITS), show promise at addressing some of these needs, but many steps still remain between theory and practice.

Therefore, constructing new tools which can allow policy makers to correctly evaluate novel technologies, such as real-time information provision or dynamic congestion pricing, is of the utmost importance. Simultaneously, there remain many opportunities to improve modeling fidelity with regard to user and system behavior. Although models are only one component of the transportation planning process, providing the best quantitative assessments possible is essential for properly choosing the best set of alternatives for implementation.

The focus of this dissertation is on network models, which represent large-scale transportation systems, typically on the scale of a major metropolitan area, although larger regional models can be constructed as well. These systems are represented mathematically by a set of nodes, connected by another set of arcs, and by various properties associated with these nodes and arcs. For instance, nodes may represent physical intersections, and arcs the roadways connecting them, with congestion-related delays associated with the arcs, and signal delays associated with the nodes. More abstract formulations are possible: arcs can represent transit lines connecting stops, or other modes of transportation such as air or sea travel; arcs can express transferring from one mode of transit to another; nodes can represent concentrated producers and attractors of travel demand; and so forth. The results of these models can provide policymakers with information such as congestion levels, toll revenues, air quality, and environmental justice.

The ultimate goal is *policy evaluation* or alternatives analysis: if a new freeway is constructed, what will be the impacts on delay, emissions, and so forth? If an innovative technology such as dynamic congestion pricing or real-time travel information is implemented, what benefits can be seen? If transit service is upgraded, what increase in ridership will be seen? Can the models themselves suggest options policy makers have not yet considered? All of these questions depend fundamentally on the ability of network models to represent users' responses to these policies. An intermediate goal, then, is to represent the *collective behavior* of these users under both prevailing system conditions, and under those obtaining under the proposed alternatives. It should be clear that this collective behavior is rooted in *individual behavior*, which must itself be properly understood before a rigorous model of collective behavior can be constructed, and therefore before the construction of rigorous policy evaluation models as well.

Many such models have been constructed over the past six decades of transportation research. However, one aspect of transportation systems which has yet to be fully integrated into network models is *uncertainty*, namely, the inability of both users and planners to accurately predict system conditions on any given day, even in the short run (due to traffic accidents, late buses or trains, or poor weather) and especially in the long run (due to forecasting errors). The aim of this dissertation is to provide a comprehensive and systematic study of operational, supply-side uncertainty in transportation networks, defined more precisely in the following sections, from their causes to their effects on individual and collective behavior, and finally to the impacts on policies aimed at improving transportation systems.

1.2 Motivation

Uncertainty pervades transportation systems, and plays an important role in nearly every aspect of the field. In operations and planning, both the causes and effects of unreliable travel times, network disruptions, and demand variations are interesting and useful topics, encompassing incident modeling, the effect of weather on driver behavior and roadway capacity, risk aversion, and the impact of reliability on mode choice, route choice, and logistics, to name just a few. It is impossible to completely remove uncertainty from the planning process: since transportation systems are inherently complex, interconnected with other systems, and driven by human behavior, future conditions cannot be specified with any degree of accuracy. Even in the short term, it is very difficult to predict the level of congestion on a given day because of weather or unpredictable events such as traffic accidents. However, while predicting the future exactly is practically impossible, important theoretical and practical benefits can be obtained by directly addressing uncertainty in the modeling process, which is the topic of this dissertation.

Furthermore, one cannot simply wish away this uncertainty by replacing stochastic parameters with their mean, or other deterministic equivalents. In the context of long-term demand uncertainty, Waller et al. (2001) show that using the mean future demand systematically underestimates expected travel times, as compared to the true expectation based on the probability distribution for demand. This is not an artifact of this particular context, and is a general property of stochastic modeling. In fact, for some stochastic programs, the difference between the expected performance of a deterministic model and the true optimal objective function can be arbitrarily large (Wallace, 2000; Higle and Wallace, 2003). Lium et al. (2009) provide a particularly eloquent explanation of this effect.

"Uncertainty" is a very broad term. To narrow the scope appropriately, one possible classification of uncertainty sources in the planning and operations domains is shown in Table 1.1. These are divided according to the time scale on which they operate, whether on the short-term, day-to-day level, or in the long-term, multi-year scale, as well as to their impact on either demand for travel, or supply provided by transportation infrastructure. Note that this partitioning is not exclusive; events such as poor weather impact both demand for travel as well as capacity offered by roadways.

The focus in this dissertation is on operational, supply-side uncertainty — that is, factors such as incidents and poor weather will be considered, but not difficulties in forecasting long-range travel demand. Still, one theme which will become evident is the *need to consider operational uncertainty even when*

Table 1.1: Taxonomy describing sources of uncertainty in transportation systems, with examples.

	Short-term (operational)	Long-term (planning)
Supply-side	Incidents Poor weather	Advances in ITS Future infrastructure
Demand-side	Daily demand fluctuation Poor weather	Future land use Long-term economic growth

conducting long-term planning. Day-to-day events have significant impacts on habitual mode choice and route choice decisions made by travelers: for instance, the possibility of missing a transfer discourages use of public transportation, and risk-averse travelers choose routes which are slightly longer on average, but which offer greater reliability. It is impossible to correctly characterize these long-term behaviors without accounting for the short-term uncertainty which shapes these decisions.

Closely related to operational uncertainty is real-time information provision to travelers, since this information only provides value because system conditions cannot be accurately predicted in advance. Owing to its real-time nature, many forms of information are available to travelers already *en route*, such as in-vehicle navigation systems, variable message signs (VMS) informing drivers of incidents or travel times, and radio traffic reports. This offers drivers the opportunity to change their route in response to this information. Furthermore, if information is reliably provided, travelers' decisions may be altered from the beginning, in anticipation of receiving additional information at a later time.

One other important effect of uncertainty considered in this dissertation is its effect on decisions made by risk-averse travelers. For instance, the consequences of late arrival when commuting to work can be substantial, and cannot be “balanced” by an occasional early arrival. In such environments, it is reasonable to expect that some drivers will choose routes which are longer on average, but whose travel times are more reliable. Recent research suggests that this effect is significant, and is important to consider when modeling travel decisions.

1.3 Problem Statements

As suggested in the previous sections, three main types of problems are considered in this dissertation: describing individual behavior, describing collective behavior, choosing network improvements optimally. This section defines these more precisely, along with their exact scope, and our assumptions regarding system and user behavior. The following section describes the mathematical notation which will be used to represent these quantities throughout.

In particular, operational, supply-side uncertainty is represented by assuming that each arc in the network can exist in one of several discrete *states* according to a stationary probability distribution which does not depend on users’ actions. Each state must directly influence the delay users experience on an arc, although this delay may still depend on the number of other users on the arc to represent congestion. For example, on an arc representing a highway, different states might represent normal operating conditions, a minor incident, a severe incident, and poor weather conditions. Arcs representing turning movements at an intersection might have states representing red and green signal indications, while arcs representing transfers at a bus stop might have states representing a "typical" transfer, and one in which a bus is overcrowded and one must wait for the next. Many other interpretations are possible, and this dissertation treats these states in a general manner which does not require one interpretation over another. Multiple interpretations may be also freely mixed as far as the following assumptions are judged representative of the system being modeled.

Assumption 1. *Users know and accurately perceive all the relevant characteristics for each state (e.g., congestion functions for highways, expected delay for red indications at signals, expected time to wait for the next bus), either through firsthand experience gathered over repeated travels, or through technology based on archived travel data.*

Assumption 2. *Users also know and accurately perceive the probability that each state occurs.*

Assumption 3. *The state that any given user experiences on an arc is independent of the state of any other arc that user experienced, and independent of any states that other users have experienced (even on the same arc).*

Assumption 3 warrants further explanation. In particular, it may seem odd that two users on the same arc may witness two different states. This is justified if the duration of certain states (e.g., the presence of an accident, or a red signal) is much shorter than the analysis period (often a peak period lasting several hours). In such cases, assuming uniform arrival of users, the number of users observing each state will be proportional to its probability of occurrence.[1]

Additionally, users may receive travel information while *en route* on their trips. In this dissertation, a specific form of local information is assumed, where drivers arriving at certain network nodes (*information nodes*) accurately learn the state of adjacent arcs. No specific technology is assumed, although a VMS is a natural example of a device providing this type of information to highway drivers. An alternate explanation, not requiring any technology at all, is that users may learn this information naturally: for instance, drivers arriving at an intersection can directly observe the state of each roadway, and adapt their choice if needed — arriving at a freeway interchange, some drivers may observe the freeway to be congested, and choose instead to remain on an arterial. Global information, such as real-time congestion updates on an in-vehicle device, can also be included in this framework if one accepts a latency or discounting effect in which information received on roadways further away is assumed to be outdated by the time the traveler arrives there.

Because users may update their travel choices *en route*, their choices cannot be described by a path in the network. Instead, their actual route in

---

[1] Alternately, assuming Poisson arrival of users, the *expected* number of users observing each state will also be proportional to its probability of occurrence.

the network will be determined by a *policy*, which maps their information at any node (including the arrival time at that node, and any message received if at an information node) to their immediate choice. Upon arriving at the next node, the policy will again be consulted and the next arc chosen, and so on, until the destination is reached. Policies are a fundamental concept in this dissertation, and are described more formally in the following section once suitable notation has been introduced.

Finally, we assume that users value trips according to the time spent traveling. This valuation can be associated with a disutility function, which need not be linear, increasing, or even continuous in arrival time. Different disutility functions represent different user behaviors, such as risk neutrality or risk aversion regarding travel delay, or the presence of a preferred arrival time as in schedule delay or deviance formulations; however, this disutility must be a function of experienced travel time only. Because of the uncertainty in arc states, this disutility is a random variable, and we take the following assumption regarding user behavior.

Assumption 4. *Users choose routing policies so as to minimize the expected disutility of their trips.*

More general definitions of disutility, depending on factors such as the smoothness of a road, or affinity or disaffinity for highways, are possible and can be accomdated in this framework with varying levels of ease; however, doing so complicates the notation considerably at the least, and is not pursued further here.

Given this conceptual background, the main problems studied in this dissertation are

Nonlinear Online Shortest Path (NL-OSP) Given fixed travel times for all arc states, determine the optimal policy for a single traveler from his or her origin to the destination, assuming that this atomic individual's choice is not significant enough to disturb these travel times. This is how individual behavior is represented.

Nonlinear User Equilibrium with Recourse (NL-UER) Given fixed demand for travel between all origins and destinations for all user classes, determine an assignment of users to policies so that no user can reduce his or her expected disutility by choosing a different policy. Note that congestion effects are considered, so each user's choice is affected by the choice of other users as well. This is how collective behavior is represented.

Information Location for Adaptive Routing Providing travelers with information can be costly, so agencies must decide the locations where information provision is most beneficial. Three variants are studied — individual information provision, when concerned only with a single individual (as when providing driving directions); uncongested information provision, when there are multiple travelers but congestion can be ignored (as in certain rural areas subject to poor weather); and congested information provision when congestion effects must be considered (as in urban areas). This is an example of a network improvement strategy which can be evaluated using the NL-OSP and NL-UER algorithms.

Congestion Pricing under Uncertainty Congestion pricing is often suggested as a tool for demand management, and encouraging users to use less congested routes during peak periods. This problem concerns how tolls should be set in the presence of uncertainty, with different assumptions about users' knowledge and the flexibility that the network manager has in dynamically varying the tolls according to the states on each arc. For simplicity, we assume a linear disutility function which allows toll costs to be directly commensurable with travel delay. This is another example of an improvement strategy which can be evaluated with the assistance of NL-UER.

Network Design Given a budget for improving arcs in some fashion (or constructing new arcs), determine an optimal set of actions to implement in order to minimize total disutility, assuming that users' collective behavior can be explained by NL-UER. This is a third example of a policy which can be evaluated using NL-UER.

Additionally, this dissertation develops a procedure to estimate the probabilities of state occurrence, and the corresponding delay functions, from ITS data.

1.4 Notation

Consider a directed, probabilistic network $G = (N, A, Z)$ consisting of a node set N, an arc set A, and a set of zones $Z \subseteq N$ representing the origins and destinations of travel. Let $\Gamma(i)$ and $\Gamma^{-1}(i)$ respectively denote the set of nodes immediately downstream and upstream of node i. Further, each arc $(i, j) \in A$ may exist in one or more discrete states $s \in S_{ij}$, each with a corresponding cost function $t_{ij}^s(x_{ij}^s)$, positive and increasing, which maps the demand x_{ij} for travel on this arc to the experienced travel delay when arc (i, j) is in state s. Let $D \subseteq Z \times Z$ represent the set of origin-destination (OD) demand pairs, and Q_{uv} the set of user classes for OD pair (u, v); as an example, these may represent risk-neutral and risk-averse travelers. For each user class, define a disutility function $f_{uv}^q(t)$ representing the burden of completing a trip whose travel time is t. Without loss of generality, let $f_{uv}^q(0) = 0$ for all travelers. Assume that the OD table $\mathbf{D}$ is completely known and deterministic, with d_{uv}^q representing the number of travelers in class q wishing to travel from node u to node v. A static viewpoint is adopted, and variations in demand and congestion throughout the study period, which is of length T, are ignored.

The probability that an individual encounters arc (i, j) in state $s \in S_{ij}$ is written as p_{ij}^s; by assumption, this probability is independent of the state of any arc previously traversed; however, the joint distribution for the states of arcs emanating from a common node may allow dependence. Let $\mathcal{S}_i = \times_{(i,j) \in A} S_{ij}$ represent the set of joint arc states at node i, with q_i^s denoting the probability that a traveler witnesses joint state $s_i \in \mathcal{S}_i$.

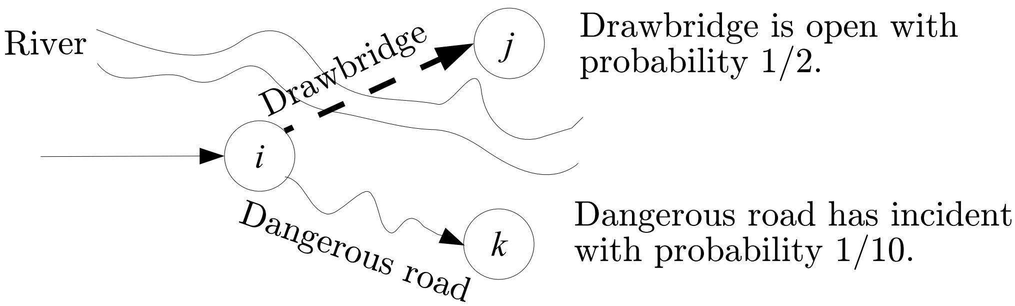

River
Drawbridge
j
Drawbridge is open with probability 1/2.
i
Dangerous road
k
Dangerous road has incident with probability 1/10.

Figure 1.1: Demonstration of information concepts.

A traveler arriving at node i may receive information on the state of adjcent arcs (i, j) (a *message*). Such a message is denoted θ_i, and is a set representing the possible states of each adjacent arc, that is, $\theta_i \subseteq S_i$. Special cases of this are *full information* on each adjacent arc, where $\theta_i \in \mathcal{S}_i$, and *no information*, where $\theta_i = \mathcal{S}_i$. Let Θ_i be the set of all possible information that can be received at node i; Θ_i must be a partition of $\mathcal{S}_i$. The probability of receiving the message θ is thus $\rho_i^\theta = \sum_{s \in \mathcal{S}_i} q_i^s$, and the probability of encountering arc (i, j) in state s conditional on message θ is expressed as

$$p_{ij}^{s\theta} = \sum_{s_i \in \mathcal{S}_i : [s_i]_j = s} \frac{q_i^s}{\rho_i^\theta} \tag{1.1}$$

For example, consider node i in Figure 1.1. Arc (i, j) is a drawbridge which can be open (O) or closed (C) with equal probability, so $S_{ij} = \{O, C\}$ and $p_{ij}^O = p_{ij}^c = 1/2$. Arc (i, k), on the other hand, is treacherous and an incident occurs with probability 1/10, so $S_{ik} = \{NI, IP\}$ where NI and IP

reflect the "no incident" and "incident present" states: $p_{ik}^{NI} = 9/10$ and $p_{ik}^{IP} = 1/10$. The set of joint states is thus

$$\mathcal{S}_i = \{s_1, s_2, s_3, s_4\} = \{[O, NI], [O, IP], [C, NI], [C, NP]\} \tag{1.2}$$

with probabilities of occurrence q_i^s shown in Table 1.2. Naturally, the rows and columns sum to the probabilities of the corresponding states. In this case, the states of (i, j) and (i, k) are independent, although this need not be the case (for instance, icy weather may make the bridge impassable and increase the risk of an accident on the dangerous road). Allowing this type of dependence is especially important when considering delay at traffic signals, as described in Chapter 2.

Now, there is a VMS at node i which is either blank, or displays one of two messages: `DRAWBRIDGE CLOSED` or `INCIDENT AHEAD`. Assuming that the VMS is always accurate (and is never blank when the bridge is closed or an incident is present), and that the `DRAWBRIDGE CLOSED` message takes precedence over the incident warning, this implies the message structure $\Theta_i = \{\theta_1, \theta_2, \theta_3\}$ where $\theta_1 = \{[O, NI]\}$ represents the blank sign, $\theta_2 = \{[O, IP]\}$ represents `INCIDENT AHEAD`, and $\theta_3 = \{[C, NI], [C, NP]\}$ represents the `DRAWBRIDGE CLOSED`. Summing the relevant entries in Table 1.2, we have $\rho_i^{\theta_1} = 9/20$, $\rho_i^{\theta_2} = 1/20$, and $\rho_i^{\theta_3} = 1/2$. For completeness, the conditional probabilities associated with these messages are shown in Table 1.3; note that messages θ_1 and θ_2 completely describe the joint state, while θ_3 leaves the state of (i, k) unknown.

Table 1.2: Joint states for demonstration of information concepts.

	NI	IP
O	9/20	1/20
C	9/20	1/20

Table 1.3: Conditional probabilities for demonstration of information concepts.

State	θ_1	θ_2	θ_3
O	1	1	0
C	0	0	1
NI	1	0	1/2
IP	0	1	1/2

Continuing with the presentation of notation and concepts, the presence of adaptive routing means that a solution to the routing problem does not consist of a single path. Rather, solutions are described as policies, which associate an arc with each node, arrival time at that node, and piece of information which can be received at that node. (This arc must be adjacent to the node.) Formally, define the set of *node states* $\Phi = \{(i, t, \theta) : i \in N, t \in T_i, \theta \in \Theta_i$. A policy π can then be defined as a function $\pi : \Phi \to A$ where $\pi(i) \in \Gamma(i)$ for all $i \in N$. Let Π_{uv} be the set of all policies connecting OD pair (u, v).

Note that a policy only prescribes the next arc to take in the path; upon arrival at the downstream end of that arc, the traveler will experience another node state, and choose the next arc according to the policy. At first glance, this approach may appear limited or myopic. In fact, the opposite is true, and a policy is a more general way to specify user behavior than a single path. If a policy prescribes the same choice of outgoing arc regardless of the information received, the traveler will follow a classical path, indicating that the set of paths is a subset of the set of policies.

Recall that each user class q has an associated disutility function $f^q(t)$ describing the preferences of each class, related to completing a trip at time t. An *optimal* policy is one which minimizes the expected disutility; denote the expected disutility of a policy π as $F(\pi)$.

As a final word on mathematical preliminaries, one should realize that if x_{ij}^s users traverse arc (i, j) in state s, they must do so in the portion of the analysis period in which (i, j) is in state s, which is of length Tp_{ij}^s. This can be addressed by using a generic delay function and scaling the demand by the proportion of time in which the link is in this state, that is, by calculating the delay $f(x_{ij}^s/p_{ij}^s)$. Alternately, one may incorporate the probability of state occurrence directly into the delay functions. For notational simplicity we adopt the latter approach, but this factor must be kept in mind when specifying delay functions.

1.5 Contributions

This dissertation represents the first attempt to model the effects of operational, supply-side uncertainty in a comprehensive manner, starting with analysis of travel data to relate the probabilities of different states to roadway geometry and other factors, using these probabilities to represent the collective behavior of users who may learn information and adapt their travel choices *en route* and exhibit sophisticated preferences in the face of this uncertainty, and finally evaluating a variety of improvement strategies which can improve the performance of the network.

While previous researchers have studied individual aspects of this problem, as detailed in the literature review sections of the succeeding chapters, the presence of a unified framework carries both theoretical and practical advantages. From the perspective of rigor, using a common set of assumptions helps ensure consistency between the component models regarding how uncertain conditions, user behavior, and other elements, are modeled. From the perspective of practice, data collection needs are minimized since the same input data can be provided to multiple components.

Although only three specific applications of NL-UER (locating information, congestion pricing, and network design) are investigated in detail, additional scenarios where this approach can be used should be clearly evident. Air quality assessment, measuring and quantifying environmental justice, and comparison of different mass transit options can all be undertaken with the appropriate methodological development and integration with existing models, to name only a few.

This work also extends earlier research in network equilibrium under uncertainty; for instance, in the course of developing algorithms for NL-UER, more efficient methods for the linear version of UER were discovered as well, as were procedures for avoiding problems with cyclic networks. Likewise, this dissertation provides a faster algorithm for solving NL-OSP than that previously developed by the author.

Finally, as with many other network problems, applications in other fields involving congestible and uncertain systems (such as power distribution or communication networks) can be imagined by drawing appropriate parallels.

1.6 Organization

The goal of this dissertation is to develop models tracing the impact of operational, supply-side uncertainty from its sources, to its macroscopic effects on user choices and network conditions, allowing planners to conduct more refined analyses. The dissertation can broadly be classified into three parts. First, a method for quantifying travel time distributions is provided, accounting for phenomena such as incidents and weather. Second, algorithms are presented to identify the corresponding effect on individual routing behavior and macroscopic system conditions at equilibrium. Finally, three diverse applications of such models are presented, along with additional methodological development where needed: deciding where to provide information, congestion pricing, and infrastructure investments such as facility expansion or construction.

A more detailed outline of the remainder of the dissertation is as follows:

Chapter 2. Determining Travel Time Distributions Before the effect of uncertainty on travel decisions can be considered, variability in travel delay must first be quantified. ITS can provide empirical distributions on some segments, but coverage is not universal and methods of estimating these distributions on other segments must be developed as well.

Chapter 3. Routing and Equilibrium Given delay distributions, the effect of uncertainty, risk preferences, information provision, and adaptive routing on individual behavior is described. Next, a macroscopic view is provided of a system equilibrium, where the travel times are partially endogenous and determined by collective routing decisions. Multiple user classes are considered.

Chapter 4. Analysis of Routing and Equilibrium Algorithms The practical performance of the routing and equilibrium algorithms must be studied, in order to ensure reasonable computation time. Furthermore, sensitivity of user behavior to network structure as well as the severity and likelihood of disruptive events will be considered.

Chapter 5. Mitigation Strategies This chapter considers three strategies for decreasing the burden of travel in uncertain networks with adaptive routing. In particular:

Information Provision Providing information at *every* node is often infeasible; in fact, when choosing where to locate devices such as VMSs, information is actually provided at very few nodes. Using the equilibrium framework from the previous chapter, the optimal locations for such devices can be determined in a quantitative and rigorous manner.

Congestion Pricing Economists have long suggested that appropriate tolls can remove congestion externalities associated with travel and lead to improvements in system conditions. Methods for applying pricing in uncertain environments are described, including scenarios in which the network manager must levy the same tolls each day, and in which the tolls can be varied flexibly in response to observed conditions.

Network Design Alternative analysis is the canonical application of planning models, and methods for determining optimal network expansion or enhancement options are described, again accounting for uncertainty, information provision, risk preferences, and adaptive routing.

Chapter 6. Conclusion The key contributions are summarized, and directions for future research are identified.

The logical relationships between these chapters, and the real-world phenomena they model, are shown schematically in Figure 1.2. Within each chapter, literature relevant to the particular topic will be thoroughly reviewed, and the models are applied to an example networks of varying sizes.

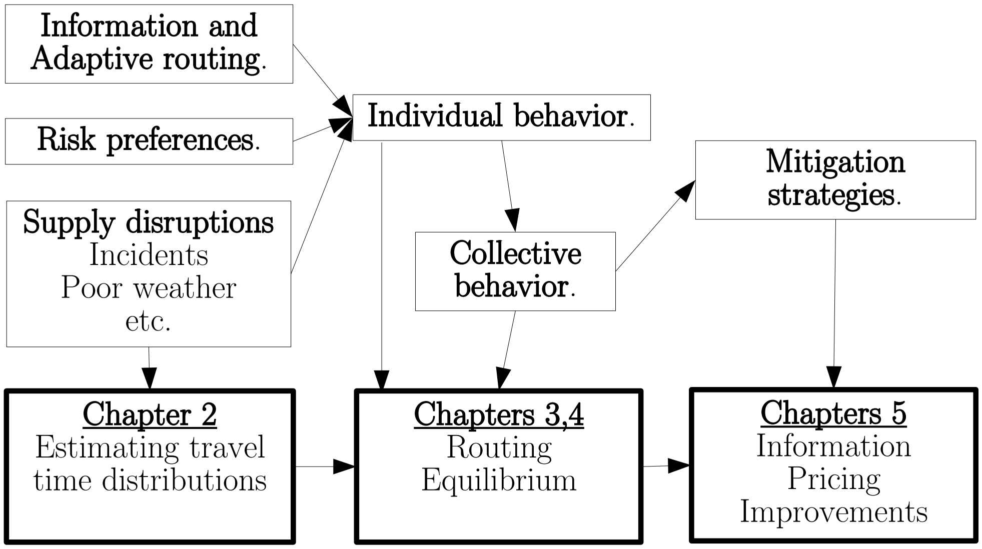


Figure 1.2: Schematic of dissertation components.

Chapter 2

Causes: Determining Travel Time Distributions

2.1 Introduction

For some time, researchers and practitioners have been aware of the costs of uncertain travel, and have begun to adapt their methods accordingly. For example, in logistics, adaptive and stochastic shortest path algorithms (Polychronopoulos and Tsitsiklis, 1996; Miller-Hooks, 2001; Waller and Ziliaskopoulos, 2002; Gao, 2005) allow vehicle routes to updated in response to travel information. In transportation planning, incorporating the value of travel reliability has been found to significantly enhance mode choice models (Small et al., 2005; Pinjari and Bhat, 2006; Liu et al., 2007). From the perspective of adaptive congestion pricing, properly accounting for uncertain conditions is needed to ensure optimum conditions obtain (Kobayashi and Do, 2005; Lindsey, 2008).

All of these models rely on facility-level descriptions of uncertain travel, often requiring an explicit probability distribution for travel time as an input. In some locations, traffic detectors record and archive speed data which may be used to generate these distributions empirically. However, coverage is often sparse, commonly existing only on major freeways in metropolitan areas (Lomax et al., 2003). As a result, while such data is invaluable for beginning to study facility reliability, additional modeling is needed to estimate distributions for *all* roadway segments in a region, freeway and arterial.

In general, one is concerned with the distribution of one or more *operational metrics*, defined as any physical quantity describing traffic flow on a single roadway segment. This includes quantities such as capacity, free-flow speed, average speed, travel time, queueing delay, average capacity reduction due to an incident, and heavy-vehicle proportion, while excluding measures such as total system travel time (which describes the entire network, not a single segment), roadway geometry (which describes the facility itself, not traffic flow), and the average annual number of icy days (which describes the region). Although the latter two have an impact on operational measures, and are included as part of the analysis, they are not the principal quantities of interest.

The main contribution of this chapter is the development of a statistical method to generate probability distributions for operational metrics, even where no archived data is available, based on facility-specific attributes. One specific application is generation of state-dependent delay functions $t_{ij}^s(x_{ij}^s)$ for use in the algorithms presented in the succeeding chapters. In the demonstrations at the end of the chapter, these methods attempt to capture the specific effect of incidents and weather on travel times, although the method can account for other factors as well. Although variation in travel demand is not part of the stated scope of this dissertation, it nevertheless plays a major role in travel time distributions. Furthermore, its effect is present in any observed data set, unless one performs additional steps which reduce the sample size substantially. Therefore, this chapter also includes some discussion of how demand variations can be included in an analysis of travel time uncertainty.

The remainder of this chapter is organized as follows. Section 2.2 describes past research into quantitative descriptions of uncertainty on roadways, and Section 2.3 describes the method used to estimate distributions for operational metrics in the absence of data, along with procedures for representing signal delay on arterials with state-dependent delay functions. Supply-side uncertainty is treated first, and then discussion is provided on using demand-side uncertainty to refine the estimates. Section 2.4 demonstrates these methods on two data sets, one from Dallas, Texas, the other from Seattle, Washington. Finally, Section 2.5 concludes the chapter and summarizes the key contributions.

2.2 Literature Review

The research literature contains relatively little on constructing complete probability distributions for travel time or other statistics, often emphasizing estimation of statistics such as confidence intervals or the proportion of late trips (Lomax et al., 2003), or quantifying the proportion of delay attributed to nonrecurring causes (Skabardonis et al., 2003; Lindley, 1987). Still, considerable research has been conducted on several related problems.

Predicting the impact of incidents requires two distinct efforts: estimating the effects of incidents that occur, and estimating the likelihood of incidents in the first place. Regarding the former, researchers have employed analytical approaches based in traffic flow theory (Wirasinghe, 1978; Morales, 1986; Boyles and Waller, 2007c), as well as statistical approaches based on field data (Golob et al., 1987; Garib et al., 1997). These are often coupled with models predicting the duration of incidents, for which a number of statistical techniques have been applied, including linear regression (Garib et al., 1997), Poisson regression (Jones et al., 1991), nonparametric regression (Smith and Smith, 2002), hazard-based models (Nam and Mannering, 2000), decision trees (Ozbay and Kachroo, 1999), and Bayesian methods (Ozbay and Noyan, 2006; Boyles and Waller, 2007c). Multiple techniques also exist for estimating incident frequency, often as functions of roadway geometry, weather, and flow (Karlaftis and Golias, 2002; Golob and Recker, 2003).

The Highway Capacity Manual (Transportation Research Board, 2000) provides some guidance on the impact of poor weather, suggesting reductions in both capacity and free flow speed as a result of rain, snow, or fog, based on previous research into these factors (Lamm et al., 1990; Ibrahim and Hall, 1994; Hogema et al., 1994; Aron et al., 1994). This information can be combined with regional historical weather data to estimate both the frequency of these events, as well as their impact on the transportation system.

Researchers have also considered the impact of demand fluctuations; almost by necessity, these approaches are macroscopic in nature, rather than facility-specific. The effect of day-to-day demand variations has been studied using simulation techniques (Asakura and Kashiwadani, 1991), equilibrium sensitivity analysis (Bell et al., 1999), and statistical techniques (Clark and Watling, 2005), providing some initial insight on how to model this phenomenon.

This chapter builds on these works by connecting these individual causes of uncertainty to probability distributions for operational metrics, potentially accounting for multiple causes at once. Additional methodological contributions are found in the derivation of average signal delay conditional on the signal indication upon arrival, and in the analysis of demand uncertainty *vis-à-vis* the distributions of operational metrics.

2.3 Method

The main contribution of this chapter is the development of techniques for quantifying operational uncertainty on highways, presented here in three subsections. First, a procedure for estimating distributions of operational metrics on freeways is presented in Section 2.3.1. Section 2.3.2 complements this by deriving analytical expressions for signal delay on arterials, using queueing theory to estimate delays conditional on arrival at different signal indications. Lastly, Section 2.3.3 considers how uncertainty in travel demand affects travel speed distributions, adopting a macroscopic approach using arc delay functions.

2.3.1 Supply-Side Uncertainty on Freeways

Many urban areas routinely collect detailed operational traffic data on freeways, using induction loop detectors, side-fire radar, and other technologies; Waller et al. (2008) provides an extensive review of data collection and archival methods. For arcs where this data is available, an empirical distribution of travel time, speed, volume, and other measures is directly available. When combined with data indicating the "state" of the freeway (e.g., the presence of an incident or poor weather), one can then estimate state-dependent delay functions. This section describes a procedure for using such data to estimate distributions of operational metrics where direct observations are not available. While these distributions can be applied in multiple ways, in the context of this dissertation they are most useful for estimating state-dependent delay functions for freeway arcs where no data is present.

Briefly, for a given operational metric, the procedure uses the available data to identify which family of probability distribution (e.g., normal, gamma) best describes the observed variation that metric within each state, and uses a regression model to relate these distributions' location and shape parameters to roadway characteristics, such as geometry and position within the network. These regression models can then be used to estimate the parameters for these distributions on any freeway arc.

The first step is to identify the relevant states that freeway arcs can exist in. The appropriate state definition depends primarily on data availability (one must know the prevailing state for each travel data observation), but also on modeling scope, geographic resolution, and other factors. For concreteness, in this section we consider four possible states — “no incident, good weather” (NIGW), “no incident, poor weather” (NIPW), “incident present, good weather” (IPGW) and “incident present, poor weather” (IPPW) — although the procedure is certainly not restricted to this configuration.

Concurrently, one should specify the operational metric of choice. This metric should be readily calculable from the available data. If the metric is directly observed, as is often the case with traffic volume or speed, each measurement can be treated as a separate observation. Other metrics may require additional filtering of the data set: if one is interested in free-flow speed, the data set should be restricted to observations with very low volume. If one is interested in the dependence of roadway capacity on freeway state, one option is to consider the highest recorded volume during each state, at each detector location, or a high percentile if outliers are a concern.[1]

At this time, it is also appropriate to consider the roadway characteristics which will be used as explanatory variables for the regression. Potential factors to include are lane width, shoulder width, number of lanes, interchange spacing, lane position of the detector, distance from the city center, proximity to weaving sections or other bottlenecks, speed limits, roadway curvature, roadway grade, peak hour factors, the proportion of heavy vehicle traffic, and

---

[1] More sophisticated procedures are clearly possible. One alternative, based on the fundamental relation $q = uk$ between volume q, space-mean speed u, and density k, is to consider clusters of neighboring observations and seek points where $dq/dk = k(du/dk) + u$ nearly vanishes, i.e., where $du/dk \approx -u/k$

any other roadway characteristic that could influence its reliability. Clearly, the exact set of factors which will be used depend on data availability, and the factors deemed most significant in a particular region.

Next, the necessary data sets must be assembled, including the travel data and the operational metrics, which must be partitioned according to the freeway state at the time of measurement. With the state definitions used in this example, incident logs and historical weather records must be consulted in order to classify observations of the operational metric into these categories.

The best-fitting distributions are then identified using the following procedure: let $\mathbb{P}$ be the set of candidate probability distribution families, and A_D the set of arcs with detector data. Then, for each arc $a \in A_D$, for each state $s \in S_a$, and each distribution family $\mathcal{P} \in \mathbb{P}$, identify the distribution parameters maximizing the likelihood of the observed sample. A chi-squared statistic $(\chi^2)^{\mathcal{P}}_{as}$ can then be calculated, representing the goodness-of-fit for this distribution.

These are used to generate a numerical ranking $R^{\mathcal{P}}_{as}$ of the distributions, where the lowest rank is associated with the best-fitting distribution, that is, the lowest $(\chi^2)^{\mathcal{P}}_{as}$. A rank-sum $R^{\mathcal{P}}_s = \sum_{a \in A_D} R^{\mathcal{P}}_{as}$ is then calculated for each distribution and each state; the best-fitting distribution for each state is the one with the lowest rank-sum.

Equipped with the best-fitting distributions for each freeway state, the likelihood-maximizing parameters from each detector location are then retrieved — for example, this would include the mean and the variance for the normal distribution, and two shape parameters for the gamma distribution — and regressed against the roadway characteristics to identify the relationship between these and the travel time distributions.

These regression models can then be applied to all freeway arcs (not just those in A_D), using their physical characteristics to estimate the distributions within each state. For the algorithms presented in succeeding chapters of this dissertation, it is necessary to estimate state-dependent delay functions $t^s_{ij}(x^s_{ij})$ as a function of roadway flow. One approach is to choose a general form for the delay function, and obtain its parameters by repeated application of the above procedure with different operational metrics. For instance, if one is using the canonical Bureau of Public Roads (BPR) formula

$$t^s_{ij}(x^s_{ij}) = (t_f)^s_{ij} \left(1 + \alpha \left(\frac{x^s_{ij}}{c^s_{ij}} \right)^{\beta} \right) \tag{2.1}$$

the state-dependent free-flow travel time $(t_f)^s_{ij}$ and capacity c^s_{ij} can be selected as the operational metrics.

Last but not least, the probability that each state occurs must also be calculated. In the case of weather, simple consultation of historical observations should suffice, with the probabilities more or less constant across the network. For incidents, it may be desirable to apply one of the models developed in the previous literature (Section 2.2) which relate incident frequency to roadway geometry and other factors, and incident probabilities calculated individually on different arcs. In general, depending on the state definition, an additional regression model may be needed to relate the state probabilities to roadway characteristics.

2.3.2 Supply-Side Uncertainty on Arterials

While the procedure in Section 2.3.1 can be used to estimate delay functions and facility reliability on freeways, arterial streets function in a fundamentally different manner which requires a different modeling approach. In particular, in most urban settings, arterial congestion is primarily determined by traffic signals, rather than by bottlenecks or oversaturation at intermediate points. This suggests that a proper representation of arterial traffic congestion requires modeling signals with some degree of realism.

This has typically been avoided in regional network modeling, due to difficulties in obtaining signal timing plans, in differentiating between different turning movements, and in accurately representing a process where different vehicles on the same arc experience radically different delays within a static modeling framework. One approach is to expand signalized intersections to include each turning movement explicitly (Figure 2.1). Then, the expected delay for each lane group can be related to the travel demand, saturation flow, and green time, using a formula such as

$$t(x) = \frac{R^2}{2C\left[1 - \min\{1, \frac{xC}{sG}\}\right]} + 900T\left[\frac{xC}{sG} - 1 + \sqrt{\frac{xC}{sG} - 1 + \frac{8klxC}{s^2GT}}\right] \quad (2.2)$$

from the Highway Capacity Manual (Transportation Research Board, 2000), where C is the cycle length, R, G, and s the effective red time, green time, and

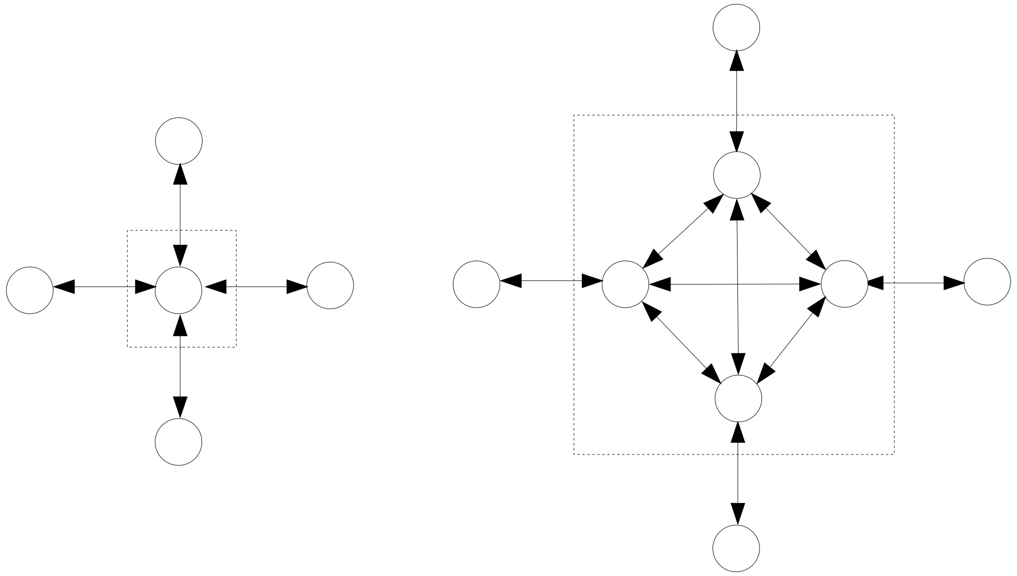


Figure 2.1: Exploding an intersection node to represent turning movements.

saturation flow for the lane group, and k and l adjustment factors to reflect intermittent oversaturation.

However, the framework of stochastic network modeling provides a new approach which can allow travelers' behavior to depend on the signal indication when they arrive at the intersection. For instance, in a grid network, there are typically many overlapping paths of nearly equal expected travel time between any two points. If a signal is red, some drivers may choose to make a right turn to avoid waiting, making an adaptive change from one path onto another to reduce their expected travel time.

The operation of traffic signals is not stochastic, actuated signals and random arrivals notwithstanding. Nevertheless, from the drivers' perspective, the signal indication upon arrival at an intersection cannot be predicted reliably unless progression is exceptionally good. This occurs because typical cycle lengths are of the same order of magnitude as the delay caused by countless factors experienced by drivers which add “noise” to travel time predictions, such as slow-moving vehicles, waiting for a gap when yielding at a turn or merge, or stopping for pedestrians to cross. Thus, we adopt a stochastic perspective to represent this uncertainty experienced by drivers, despite the deterministic operation of traffic signals.

In this setting, the delay functions must show the expected delay for travelers who arrive on green and red indication. Queueing diagrams are useful for this purpose. Consider Figure 2.2, which plots cumulative vehicle arrivals and departures for a lane group at a signalized intersection where the travel demand $x \leq sG/C$ (that is, the lane group is undersaturated). The solid line indicates the uniform arrival process, while the dashed line indicates the departure process, which is controlled by the signal indication. The horizontal distance between these curves gives the queue length at any point in time, while the vertical distance gives the delay a given vehicle will experience.

Because this lane group is undersaturated, the queue and vehicle delays are periodic according to the cycle length C so, without loss of generality, consider the first cycle interval $[0, C]$. If we define $\overline{x} = x/T$ to represent the vehicle arrival rate, the v-th vehicle will arrive at time $A(v) = v/\overline{x}$ and depart at time $D(v) = R + v/s$ if $v \leq Rs\overline{x}/(s - \overline{x})$, and at time $D(v) = v/\overline{x}$ if $Rs\overline{x}/(s - \overline{x}) \leq v \leq C\overline{x}$. (As can be readily verified, the queue clears just

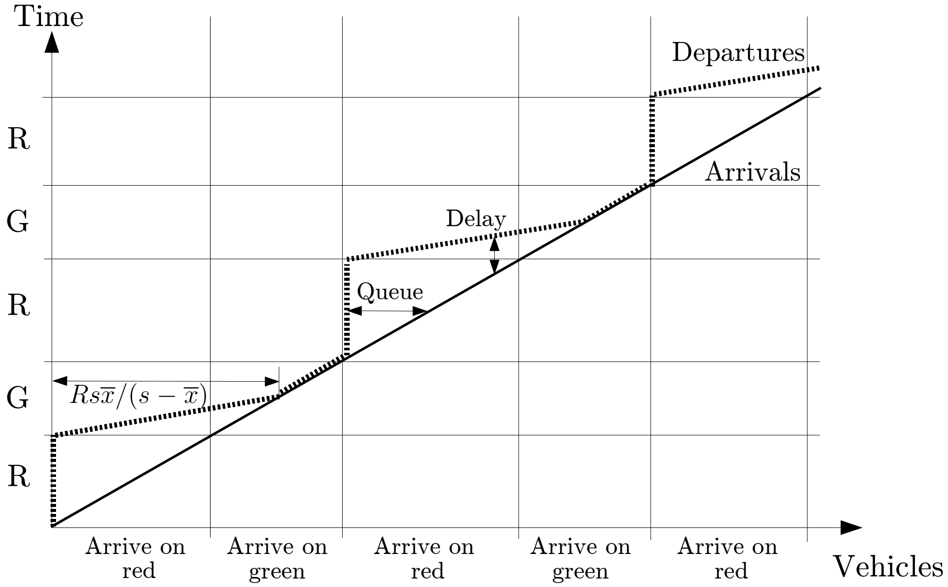


Figure 2.2: Queuing diagram for undersaturated lane groups.

as the $Rs\overline{x}/(s-\overline{x})$-th vehicle arrives; until the next red indication, arriving vehicles experience no signal delay).

Note that the arrival and departure functions $A(v)$ and $D(v)$ are piecewise linear, so for travelers arriving on red, the average delay is simply the average of $D(0) - A(0)$ and $D(\overline{x}R) - A(\overline{x}R)$:

$$t^R(x) = \frac{1}{2}\left[(R-0) + \left(R + \frac{\overline{x}R}{s} - \frac{\overline{x}R}{\overline{x}}\right)\right] \tag{2.3}$$

$$= \frac{R}{2}\left[1 + \frac{\overline{x}}{s}\right] \tag{2.4}$$

$$= \frac{R}{2}\left[1 + \frac{x}{Ts}\right] \tag{2.5}$$

For vehicles arriving on green, some will observe a queue while others will not. Those observing a queue will, on average, experience a delay of

$R\overline{x}/2s$, while the others experience zero delay. Since the number of vehicles observing a queue when arriving on green is $Rs\overline{x}/(s-\overline{x}) - R\overline{x}$, while the total number of vehicles arriving on green is $G\overline{x}$, the unconditional average delay experienced by those arriving on green is

$$\begin{aligned} t^G(x) &= \frac{R\overline{x}}{2s}\frac{Rs\overline{x}/(s-\overline{x}) - R\overline{x}}{G\overline{x}} && (2.6) \\ &= \frac{R\overline{x}}{2s}\frac{R}{G}\left[\frac{\overline{x}}{s-\overline{x}}\right] && (2.7) \\ &= \frac{Rx}{2Ts}\frac{R}{G}\left[\frac{x}{Ts-x}\right] && (2.8) \end{aligned}$$

The oversaturated case $x \geq sG/C$ is more difficult, because the queue behavior is no longer periodic. Instead, its average size will increase over the duration of the analysis period (Figure 2.3). On the other hand, it is easier to write an equation for the departure curve

$$D(v) = \frac{v}{s} + R\left(\left\lfloor \frac{V}{sG} \right\rfloor + 1\right) \tag{2.9}$$

using the floor function $\lfloor x \rfloor = \max\{z \in \mathbb{Z} : z \leq x\}$. Thus, the delay experienced by the v-th vehicle is

$$D(v) - A(v) = \frac{v}{s} - \frac{v}{\overline{x}} + R\left(\left\lfloor \frac{V}{sG} \right\rfloor + 1\right) \tag{2.10}$$

Assuming no queue at the start of the analysis period, and that $t = 0$ corresponds to the start of the red indication, the average delay experienced by those arriving on red or green can be found through integration. Let $n_C = \lceil T/C \rceil$ represent the number of cycles (including partial ones) which occur

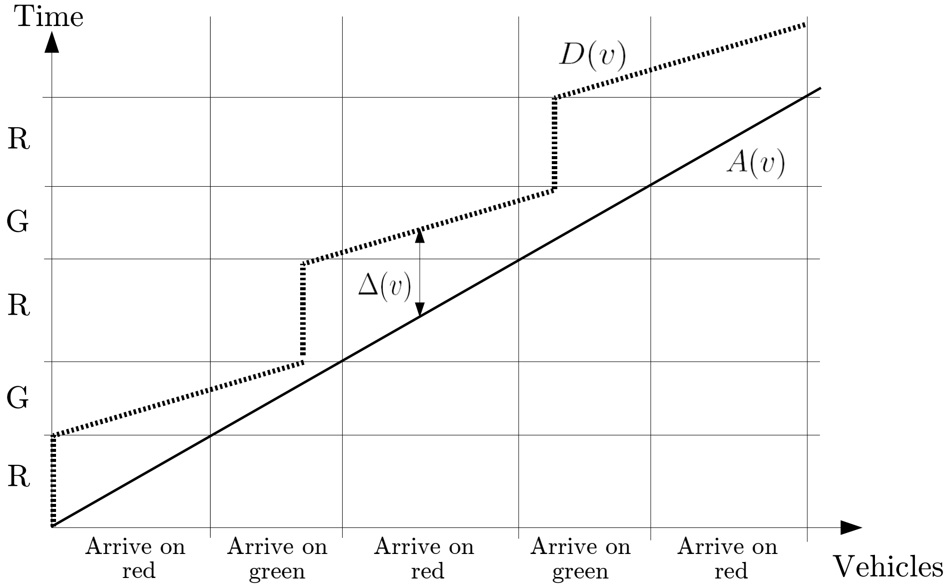


Figure 2.3: Queuing diagram for oversaturated lane groups.

during the analysis period. Then

$$t^R(x) = \frac{\displaystyle\sum_{i=1}^{n_C} \int_{C(i-1)x/T}^{\min\{x,(C(i-1)+R)x/T\}} \left[\frac{v}{s} - \frac{vT}{x} + R\left(\left\lfloor \frac{V}{sG} \right\rfloor + 1\right)\right] dv}{\displaystyle\sum_{i=1}^{n_C} \min\left\{\frac{xR}{T}, \left(1 - \frac{C(i-1)}{T}\right) x\right\}} \tag{2.11}$$

$$t^G(x) = \frac{\displaystyle\sum_{i=1}^{n_C} \int_{\min\{x,[Cx(i-1)+G]/T\}}^{\min\{x,Cix/T\}} \left[\frac{v}{s} - \frac{vT}{x} + R\left(\left\lfloor \frac{V}{sG} \right\rfloor + 1\right)\right] dv}{\displaystyle\sum_{i=1}^{n_C} \min\left\{\frac{xG}{T}, \left[1 - \frac{Ci}{T} - \frac{R}{T}\right]^+ x\right\}} \tag{2.12}$$

where $[\cdot]^+ = \max\{\cdot, 0\}$.

2.3.3 Demand-Side Uncertainty

Although this dissertation's scope does not explicitly consider demand-side uncertainty, demand variations still play a significant role in determining distributions of operational metrics. It also provides another avenue of approach for considering distributions on arterials, where ITS coverage is often poor. Nevertheless, since demand data are harder to obtain than speed observations, and since the primary focus of this dissertation is on supply-side uncertainty, the procedures developed in this subsection should be viewed as an enhancement or refinement of the procedures described above for supply-side uncertainty. The initial focus of the subsection is on demand-side uncertainty alone, before integrating these results with those for supply-side uncertainty.

Specifically, a procedure is developed to show how demand variability affects travel speeds. Since demand is inherently macroscopic in nature, demand variations must be considered at the network level, rather than the facility or corridor level. Other operational metrics can be treated with a similar derivation.

The degree to which arcs are affected by demand uncertainty depends on two primary factors. First, the "typical"operating condition must be considered, as arcs with either low congestion or high congestion will be less affected by fluctuations in demand, an effect we term *intrinsic sensitivity.* Second, the network structure must be considered: where alternate routes exist, arcs are less sensitive to demand fluctuations than where there is no viable alternative; this effect we term *extrinsic sensitivity.* Each of these is discussed in turn, and then combined into a single measure. The resulting formulas will allow the mean and variance of travel speed to be estimated, incorporating demand uncertainty.

Since we are concerned with travel speeds, and since cost functions are generally expressed in terms of travel time, we apply the transformation $s = L/t$, with s the travel speed, L the arc length, and t the traversal time. The sensitivity of an arc's speed to a change in demand can be represented by the derivative

$$\frac{ds}{dx} = -\frac{L}{t^2}\frac{dt}{dx} = -\frac{s^2}{L}\frac{dt}{dx} \tag{2.13}$$

For instance, using the BPR function (2.1) produces

$$\frac{ds}{dx} = -\frac{t_0 \alpha \beta}{L c^\beta} x^{\beta-1} s^2 \tag{2.14}$$

indicating that this arc is "robust" to demand uncertainty if either the demand d or travel speed s is low. That is, changes in demand have less effect if few people are using the arc (close to free-flow), or if the arc is already highly congested (speed cannot degrade much further), and changes in demand have the greatest effect when the quantity $x^{\beta-1}s^2$ is maximized. This can be generalized. Almost all commonly-used delay functions are increasing and convex; that is, dt/dx is positive and increasing in x, so ds/dx is small when x (and thus dt/dx) is small, or when s is small.

Quantifying extrinsic sensitivity involves relating uncertainty in macroscopic demand to the uncertainty in demand for an individual arc. Following

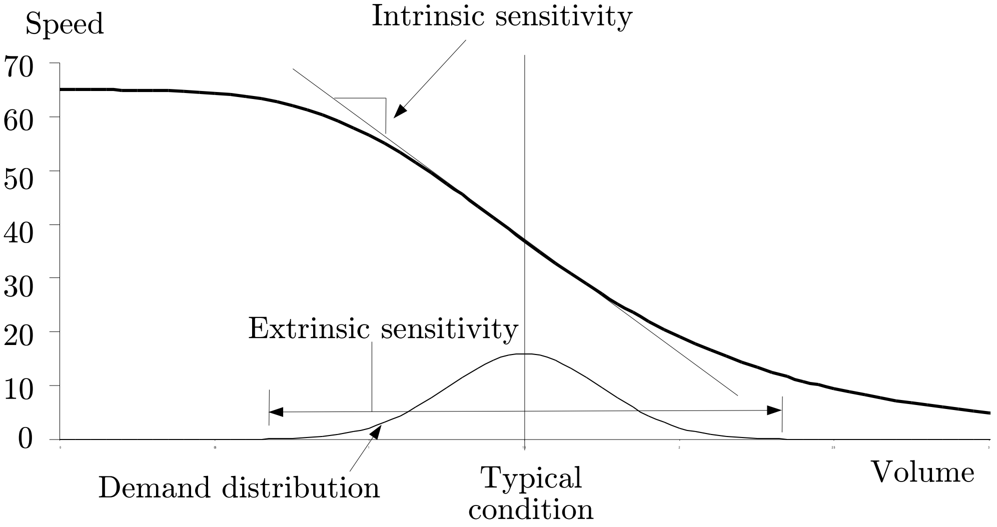


Figure 2.4: Comparison of intrinsic and extrinsic sensitivity.

Clark and Watling (2005) and Unnikrishnan (2008), let ξ^{ij}_{uv} denote the proportion of travelers from OD pair (u, v) that use arc (i, j). Then if the demand $\tilde{d}_{uv}$ is uncertain, so is the demand for travel on individual arcs, given by:

$$\tilde{x}_{ij} = \sum_{(u,v)\in D} \tilde{d}_{uv}\xi^{ij}_{uv} \tag{2.15}$$

Thus, the mean and variance of arc demand are

$$\mu_{ij} = \sum_{(u,v)\in D} E[\tilde{d}_{uv}]\xi^{ij}_{uv} \tag{2.16}$$

and

$$\sigma^2_{ij} = \sum_{(u,v)\in D}\sum_{(t,u)\in D} Cov[\tilde{d}_{uv}, \tilde{d}_{tu}]\xi^{ij}_{uv}\xi^{ij}_{tu} \tag{2.17}$$

respectively. In general it is difficult to derive the exact probability density function for arc flows, as it requires a large multiple integral ($O(n^2)$ integrations per arc). However, two special cases are worth noting:

- If OD demands are independent and normally distributed, $\tilde{x}_{ij}$ is also normally distributed with the mean and variance as given above.

- If OD demands are independent and Poisson distributed with rate parameters λ_{uv}, $\tilde{x}_{ij}$ is also Poisson distributed with rate parameter $\lambda_{ij} = \sum_{(i,j)\in D} \lambda_{uv}$.

Substituting (2.15) into (2.13) and applying the chain rule, intrinsic and extrinsic sensitivity can be combined, with the sensitivity of travel speed on arc (i,j) to a change in demand from OD pair (u,v) shown to be

$$\frac{ds_{ij}}{d(d_{uv})} = -\frac{s_{ij}^2}{L_{ij}}\frac{dt_{ij}}{dx_{ij}}\frac{dx_{ij}}{d(d_{uv})} = \frac{ds_{ij}}{dx_{ij}}\xi_{uv}^{ij} \tag{2.18}$$

Taking a tangent plane approximation to s at the point $\mathbf{D}$, the speed resulting from a change in demand $\Delta\mathbf{D}$ can be approximated by

$$s + \frac{ds_{ij}}{dx_{ij}}\sum_{(u,v)\in D}\Delta d_{uv}\xi_{uv}^{ij} \tag{2.19}$$

which, for the BPR relation, is

$$s - \frac{t_0\alpha\beta}{L_{ij}c^{\beta}}x^{\beta-1}s^2\sum_{(u,v)\in D}\Delta d_{uv}\xi_{uv}^{ij} \tag{2.20}$$

Given some density function $g_{ij}(x)$ for demand on arc (i,j), one fix a "typical" speed value s_0, and take a linear approximation at the point $x_0 = \mu_{ij}$, giving estimates of the mean and variance of travel speed as

$$E[s|s_0] \approx \int \left(s_0 + s'\left(x_0\right)\left(x - x_0\right)\right) g(x)dx = s_0 + s'(x_0)(E[x] - x_0) = s_0 \tag{2.21}$$

and

$$\begin{aligned}Var[s|s_0] &\approx \int (s_0 + s'(x_0)(x - x_0))^2 g(x)dx - (E[s|s_0])^2 \\ &= (s_0 - s'(x_0)x_0)^2 + 2(s_0 - s'(x_0)x_0)E[x] + [s'(x_0)]^2 E[x^2] - s_0^2 \\ &= [s'(x_0)]^2 Var[x] \qquad (2.22)\end{aligned}$$

with the arc subscripts omitted for brevity.

Finally, the combined effect of demand-side and supply-side uncertainty can be derived. Using the procedure in Section 2.3.1, a density function $f_{ij}(s)$ can be calculated for the speed on (i, j). Fixing x_0, the derivative $s'(x_0)$ still depends on s, implying that the above formulas condition on a given freeway operating speed. Using the law of total variance, we can write unconditional expressions for these quantities:

$$E[s] = \int E[s|s_0]f(s_0)ds_0 = \int s_0 f(s_0)ds_0 = \mu_{ij} \qquad (2.23)$$

$$\begin{aligned}Var[s] &= Var\left[E\left[s|s_0\right]\right] + E\left[Var\left[s|s_0\right]\right] = Var\left[s_0\right] + E\left[\left(s'\left(x_0\right)\right)^2 Var\left[\tilde{x}_{ij}\right]\right] \\ &= \sigma_{ij}^2 + E\left[\left(s'\left(x_0\right)\right)^2\right] \sum_{(u,v)\in D} \left(\xi_{uv}^{ij}\right)^2 Var\left[\tilde{d}_{uv}\right] \qquad (2.24)\end{aligned}$$

Again using the BPR example, this formula simplifies to

$$Var[s] = \sigma_{ij}^2 + \left(\frac{t_{ij}^0 \alpha\beta}{L_{ij} c_{ij}^\beta} x^{\beta-1}\right)^2 E[s^4] \sum_{(u,v)\in D} (q_{uv}^{ij})^2 Var[\tilde{d}_{uv}] \qquad (2.25)$$

where $E[s^4]$ is the fourth raw moment of the speed distribution found from the supply-side analysis. The same analysis can be repeated for other operational measures.

2.4 Demonstration

In this section, the procedure in Section 2.3.1 for estimating operational metrics in freeways is demonstrated through two example applications. The first application is concerned with producing probability distributions for travel speed on freeways in the Dallas-Ft. Worth metropolitan area. The second uses data from Seattle, Washington, and is concerned with estimating the capacity parameter in the BPR equation for incident and no-incident conditions.

2.4.1 Dallas

Using data obtained from the Dallas-Ft. Worth metropolitan area, the procedure in Section 2.3.1 is used to estimate probability distributions for freeway traffic speed, particularly in locations without detectors. The main approach is to estimate conditional distributions for different freeway states (incident, poor weather, normal conditions, etc.), and to produce separate regression models relating each of these distributions to the roadway characteristics. Then, for any freeway in question, the state-dependent distributions can be constructed and combined into an unconditional distribution using the law of total probability.

Three main data sets were obtained: archived loop detector observations providing speed data; a set of incident logs detailing locations, times, and durations; and a set of weather data providing information on temperature and precipitation. Loop detector data for this region is available online from a publicly available website[2]. This site provides a separate file for each day; for use in this dissertation, these were converted into separate files for each detector. Incident logs were obtained from the Texas Department of Transportation, and include a text description of each incident's location along with detection and clearance times. Daily weather data was obtained from the National Weather Forecast Office[3].

These data sets were merged into a common file, and divided into three segments: **no incident, good weather** (NIGW), **poor weather** (PW), and **incident present** (IP). As relatively few observations existed for cases where *both* an incident and poor weather conditions were present, the choice was made to omit this latter category, and classify any such observations under both the PW and IP categories. Furthermore, this data set only contained weather data at the resolution of one day, so "poor weather" was applied to all observations on days with a half inch or more of precipitation. This suffices for a demonstration; for field application, a more disaggregate data source would be highly valuable.

For each detector and each category (NIGW, PW, IP), fifteen different probability distributions were fit to the observed speed data: the normal, lognormal, beta, chi-squared, Erlang, exponential, fatigue life, Frechet, gamma, generalized extreme value, Gumbel, logistic, log-logistic, Rayleigh,

---

[2] `http://ttidallas.tamu.edu/detectordataarchive/DalTrans/Default.htm`. Accessed March 20, 2008.

[3] `http://www.srh.noaa.gov/fwd/f6.htm`. Accessed April 20, 2008.

Table 2.1: Example ranking of distributions for one detector.

Distribution	χ^2	Rank
Beta	9.826	1
Log-logistic	10.297	2
Weibull	14.62	3
FatigueăLife	14.802	4
Gamma	14.966	5
Lognormal	15.516	6
Erlang	19.212	7
Gumbel	22.481	8
Generalizedăextreme value	25.341	9
Normal	25.541	10
Frechet	31.559	11
Logistic	34.933	12
Rayleigh	42.646	13
Chi-squared	68.756	14
Exponential	119.66	15

and Weibull distributions. Based on this process, the normal distribution best describes speeds for the NIGW category, while the beta distribution best describes speeds for the PW and IP categories. Table 2.1 shows the results from this ranking process.

These distributions are specified by two parameters each: the mean μ and standard deviation σ for the normal distribution, which has density

$$f(s; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(s-\mu)^2/(2\sigma^2)} \tag{2.26}$$

and two shape parameters a and b for the beta distribution, which has density

$$f(s; a, b) = \frac{1}{B(a, b)} s^{a-1}(1-s)^{b-1} \tag{2.27}$$

where B is the beta function.

In order to estimate speed distributions at locations where no data is present, we attempt to link these distribution parameters to roadway geometry and segment characteristics using linear regression. Drawing inspiration from the Highway Capacity Manual procedure for calculating free-flow speed on freeway segments (Transportation Research Board, 2000), the independent variables chosen for regression were the lane width, shoulder width, number of lanes, interchange spacing, and lane position (that is, whether the detector is located on an inner or outer lane), while the dependent variables were the best-fitting distribution parameters for each detector.

Three such regressions were performed, for the NIGW, PW, and IP scenarios; regression results are shown in Tables 2.2 to 2.4, with t-statistics shown in parentheses. One notable result is that the R^2 values are very low for the NIGW and PW scenarios, indicating that these geometric factors do not play a significant role in determining speed distributions when no incident is present. The past literature clearly shows that geometry affects the probability of an incident occuring, and this data reveals that geometry also affects the severity of an incident (as measured by the resulting speed distributions); but this initial investigation suggests that the impact is limited beyond this.

Using this procedure, three probability distributions $f_{NIGW}(s)$, $f_{IP}(s)$, and $f_{PW}(s)$ can be constructed for any freeway segment, using its geometric properties to choose the distribution parameters. These can then be combined into an unconditional speed distribution

$$f(s) = f_{NIGW}(s)\Pr(NIGW) + f_{IP}(s)\Pr(IP) + f_{PW}(s)\Pr(PW) \tag{2.28}$$

Table 2.2: Regression results for NIGW scenario .

NORMAL	μ	σ
Constant	41.23 (10.12)	15.51 (0.45)
Outer lane dummy	—	−40.11 (−1.20)
Number of lanes	—	47.08 (1.32)
Shoulder width (m)	5.13 (3.1)	—
R^2	0.20	0.06

Table 2.3: Regression results for PW scenario.

BETA	a	b
Constant	33.07 (4.23)	25.73 (3.93)
Outer lane dummy	—	1.69 (1.52)
Number of lanes	−6.09 (−3.90)	−5.85 (−4.03)
Shoulder width (m)	2.12 (5.30)	1.79 (1.70)
Interchange spacing (mi)	−14.18 (−1.83)	−3.78 (1.50)
R^2	0.49	0.46

Table 2.4: Regression results for IP scenario.

	a	b
Constant	6.295 (2.98)	5.19 (3.51)
Outer lane dummy	—	0.90 (2.00)
Number of lanes	−0.698 (−1.20)	−0.51 (−1.10)
R^2	0.03	0.01

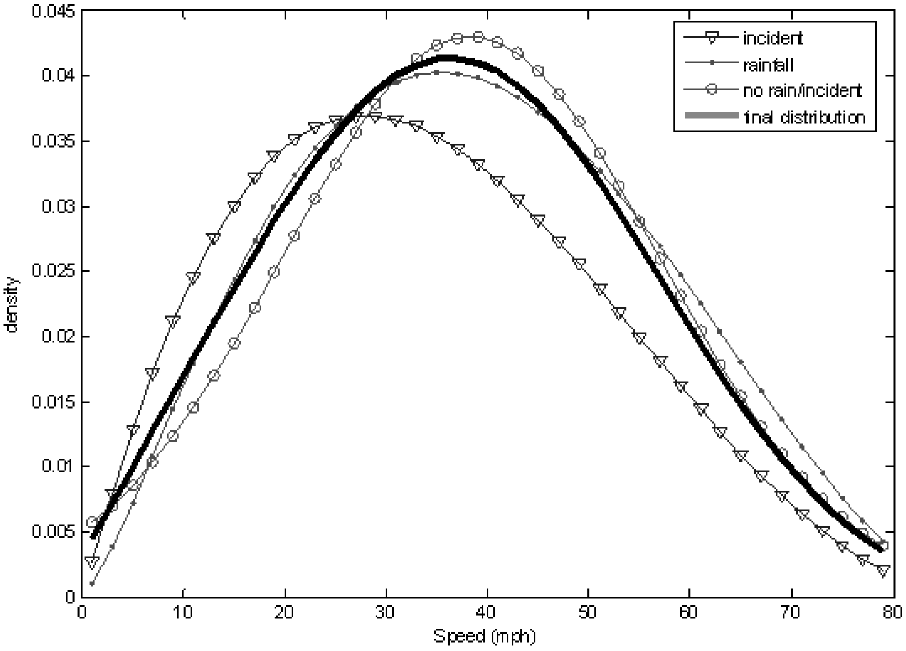


Figure 2.5: Conditional and unconditional speed distributions.

where the probability of an incident Pr(IP) can be calculated using the models mentioned in Section 2.2, the probability of poor weather Pr(PW) can be estimated from historical data, and Pr$(NIGW) \approx 1 - \Pr(IP) - \Pr(PW)$.[4] Figure 2.5 shows how these distributions are related.

2.4.2 Seattle

A second demonstration of these procedures can be applied using data from the Seattle network. Unlike the Dallas data, where the objective was

---

[4] Since a few observations are classified in both the IP and PW scenarios, the actual value of Pr$(NIGW)$ is slightly higher than this.

Table 2.5: Seattle data highway segments and network arcs.

Highway segment	Length (mi)	Network arcs
I-5 NB Seatac to Seattle	13	(7,5)
I-5 NB Seattle to WA-526	23	(5,3), (3,2), (2,1)
I-5 SB WA-526 to Seattle	23	(1,2), (2,3), (3,5)
I-5 SB Seattle to Seatac	13	(5,7)
WA-167 NB Auburn to Renton	10	(10,8)
WA-167 SB Renton to Auburn	10	(8,10)
I-405 NB Tukwila to Bellevue	14	(7,8), (8,6)
I-405 NB Bellevue to WA-524	16	(6,4), (4,2)
I-405 SB WA-524 to Bellevue	16	(2,4), (4,6)
I-405 SB Bellevue to Tukwila	14	(6,8), (8,7)

characterizing the observed speed distributions, the objective with the Seattle data is to determine delay functions $t^s_{ij}(x^s_{ij})$ which can be used in the algorithms developed in the remainder of this dissertation. This data includes travel speeds and volumes for several highway segments at the 5-minute resolution, along with an indication as to whether an incident was affecting that segment at the given time. (Considerably more information is contained in this data set, but was not used in this demonstration.)

Table 2.5 shows how the highway segments in this data set map to the Seattle network shown in Figure 2.6; this network will also be used in the following chapter to demonstrate the NL-OSP routing algorithm. Clearly, this aggregated network is a greatly simplification of the region's freeway infrastructure, and this demonstration is not intended as anything other than an example.

With this data set, each arc can exist in one of two states — no incident

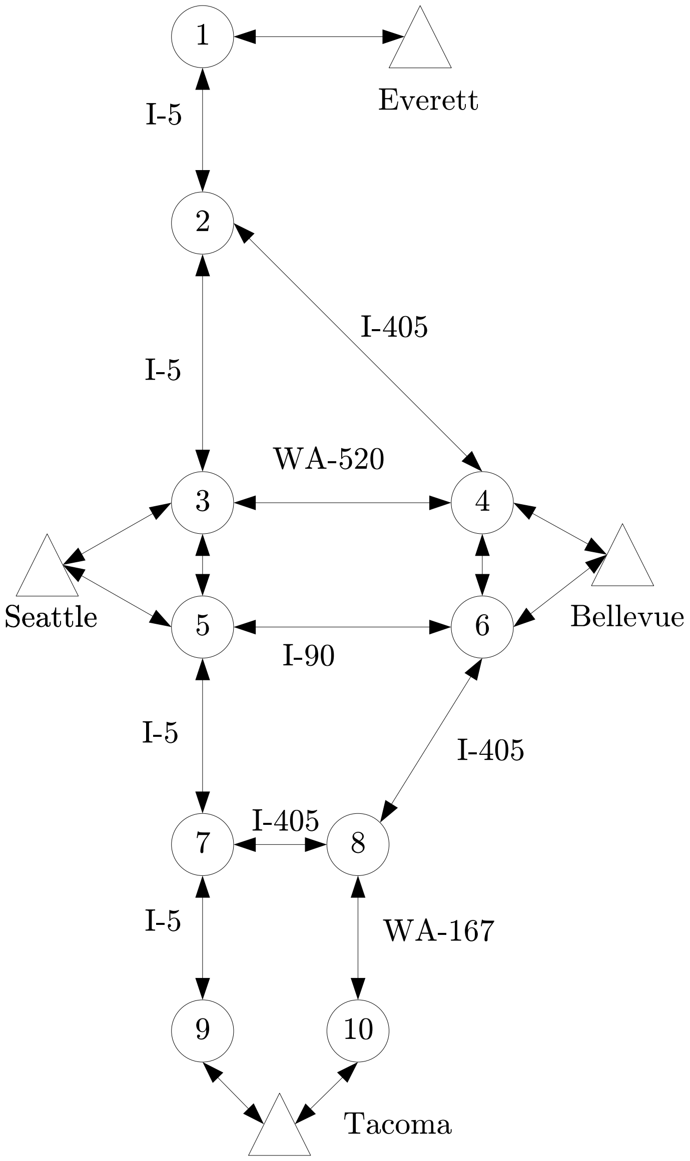


Figure 2.6: Seattle network used to demonstrate dissertation algorithms.

(NI), or incident present (IP). A delay function must be estimated for each of these, which are assumed to be of the BPR form (2.1). Thus, the free-flow travel time and capacity must be given for each arc in the network, not all of which are covered by the data set. For both the NI and IP states, the free-flow travel time are given by the segment length divided by the speed limit, converted to appropriate units[5].

Regarding capacity, recall that the "capacity" parameter c_{ij}^s in the BPR equation does *not* actually represent the true roadway capacity, but is merely a parameter used in a function converting demand for travel on an arc to the experienced travel delay, a demand which can easily exceed the true capacity at times. By contrast, the actual volume on the arc can never exceed the true capacity; this disparity is a well-known limitation of static traffic assignment models of the type used in this dissertation. For the typical BPR shape parameters $\alpha = 0.15$ and $\beta = 4$, c_{ij}^s is often taken to be the "practical capacity" of the roadway, roughly corresponding to level of service E; Kockelman (2003) estimates this to be roughly 80% of the true capacity.

If c_{ij}^{NI} represents the practical roadway capacity in the NI state, we express the practical capacity in the IP state as $\text{ת}_{ij} c_{ij}^{NI}$, where ת_{ij} represents the capacity decrease due to the incident's presence. For each arc, c_{ij}^{NI} is estimated from the data set by finding the maximum observed volume during

---

[5] The speed limit for all of the freeways in this network is 60 mph; although average travel speed is often higher than this at near free-flow conditions, the data set truncated speed measurements at the speed limit.

any 5-minute time interval, scaled to appropriate units, and multiplied by 80%.

Estimation of ח_{ij} is more involved. Simply reducing c_{ij}^{NI} to account for a lane blockage is insufficient for several reasons: the capacity at the bottleneck is not the true quantity of interest, but rather the travel delay on the entire roadway segment. Upstream of the incident, additional delay is incurred; however, downstream of the incident, some travel savings may be obtained due to a metering effect which reduces congestion below its normal level. Instead, the following procedure is used to choose ח_{ij} to match the observed decrease in travel speeds during incident conditions.

Suppressing the arc subscript for brevity, let s^{NI} and s^{IP} represent the average travel speeds during these states, and let s_f denote the free-flow travel speed. First, we determine the travel demand x leading to the travel speed s^{NI} by inverting (2.1) and expressing quantities in terms of speed:

$$x = c^{NI}\left(\frac{s_f/s^{NI} - 1}{\alpha}\right)^{1/\beta} \tag{2.29}$$

Now, we have

$$\frac{s^{IP}}{s^{NI}} = \frac{1 + \alpha(x/c^{NI})^{\beta}}{1 + \alpha(x/\text{ח}c^{NI})^{\beta}} \tag{2.30}$$

and solving for ח yields

$$\text{ח} = \frac{x}{c^{NI}}\left[\frac{1}{\alpha}\left(\frac{1 + \alpha(x/c^{NI})^{\beta}}{s^{IP}/s^{NI}} - 1\right)\right]^{-1/\beta} \tag{2.31}$$

Although it may not be immediately obvious from the above procedure, ח is in fact independent of c^{NI}, and is solely determined by s_f and the ratio s^{IP}/s^{NI}.

Table 2.6: Regression results for Seattle data.

	c	ת	$\Pr(IP)$
Constant	-66 (-0.13)	0.73 (13.2)	-0.54 (-5.08)
Distance to CBD	—	—	0.02 (6.02)
Interchange density	—	0.10 (5.93)	—
Number of lanes	1670 (12.70)	-0.01 (-1.37)	0.11 (5.93)
R^2	0.95	0.46	0.84

Lastly, we need to calculate the probability of an incident occuring, $\Pr(IP)$. This is calculated directly from the data.

Because our interest is simply in the mean values of these parameters, the distribution-fitting step can be ignored, and we proceed immediately to the regression equations used to estimate c, ת, and $\Pr(IP)$ on arcs without data. Since these statistics are calculated from all of the available data for a highway segment, only one observation exists for each segment. Given the relative paucity of data due to aggregation, only three explanatory variables were considered: the number of lanes n_L, the interchange density I_D expressed in interchanges per mile, and the distance from the Seattle city center d_{CBD}, measured in miles. After pruning insignificant variables, the resulting regression equations are

$$c \approx 1670n_L - 66 \tag{2.32}$$

$$\text{ת} \approx 0.736 + 0.142I_D - 0.015n_L \tag{2.33}$$

$$\Pr(IP) \approx 0.02d_{CBD} + 0.11n_L - 0.54 \tag{2.34}$$

Further details on the regression can be seen in Table 2.6.

Applying these equations, delay functions can be created for each arc in the Seattle network; all of the relevant parameters are shown in Table 2.7. It bears repeating that the intent of this section is not to provide a rigorous calibration of delay functions for the Seattle region, which would require a far more disaggregate and comprehensive analysis, but merely to produce an example network with some degree of *vraisemblance*.

2.5 Conclusion

This chapter considered how to quantify operational uncertainty existing in transportation networks. For freeways, a procedure was described for estimating distributions of operational metrics, allowing calculation of statistics such as the mean and variance. This allows estimation of parameters for state-dependent delay functions, which is needed for all of the methods in the remainder of this dissertation. Procedures for incorporating demand uncertainty were described, and can be implemented given availability of appropriate data. Additionally, on arterials, queueing theory was used to separately describe the expected delay for travelers arriving on red or green indications. These can be used as input to the adaptive routing and equilibrium models developed in the next chapter, and provides a new perspective on incorporating signalized intersections into network-level modeling.

Table 2.7: Seattle data highway segments and network arcs.

Arc	t_f^{NI}	c^{NI}	t_f^{IP}	c^{IP}	ת
(1,2)	7	6864	7	5229	0.762
(2,1)	7	6864	7	5229	0.762
(2,3)	14	6864	14	5229	0.762
(2,4)	16	5895	16	4200	0.712
(3,2)	14	6864	14	5229	0.762
(3,4)	10	3825	10	2903	0.759
(3,5)	4	6864	4	5229	0.762
(4,2)	16	5722	16	4451	0.778
(4,3)	10	3825	10	2903	0.759
(4,6)	3	5895	3	4200	0.712
(5,3)	4	6864	4	5229	0.762
(5,6)	10	6614	10	4858	0.735
(5,7)	13	8762	13	6847	0.781
(6,4)	3	5722	3	4451	0.778
(6,5)	10	4944	10	3704	0.749
(6,8)	9	5609	9	4487	0.800
(7,5)	13	7577	13	6009	0.793
(7,8)	2	5994	2	4848	0.809
(7,9)	12	7449	12	5523	0.741
(8,6)	9	5994	9	4848	0.809
(8,7)	2	5609	2	4487	0.800
(8,10)	10	4079	10	3251	0.797
(9,7)	12	7449	12	5523	0.741
(10,8)	10	3950	10	3304	0.836

Chapter 3

Effects: Routing and Equilibrium

3.1 Introduction

Once probability distributions are available for arc travel times, the next question is how users choose routes in such an environment. In particular, the dissertation considers how an individual user would behave (the *routing* problem), and how the collective behavior of multiple self-interested users can be described (the *equilibrium* problem).

This work is distinguished from past research in stochastic shortest paths and traffic assignment in two main ways:

- **Information provision** is considered: At certain nodes, users may learn information on the state of adjacent arcs, and are free to vary their route in response.
- **Nonlinear preferences** are considered: Users are not assumed to be simply interested in minimizing expected travel time. Rather, more sophisticated behaviors, such as risk aversion, minimizing schedule delay, or maximizing the probability of on-time arrival, can be considered.

Boyles (2006) and Boyles and Waller (2007b) considered a similar routing problem; the dissertation builds on this work by providing a more efficient algorithm, by accounting for a broader class of user preferences, and by considering a different information provision scenario. Unnikrishnan (2008) considered a similar equilibrium problem, but with only linear preferences, and restricted to application in acyclic network. This dissertation extends this research to account for cycles, develops a more efficient solution method, and is able to account for nonlinear preferences.

This chapter first discusses the routing problem in Section 3.2, followed by the equilibrium problem in Section 3.3, and a summary in Section 3.4. Each section provides a fuller outline of its content, a review of relevant literature, discussion of the problem at hand, and exact solution algorithms.

3.2 Routing

Recall that the routing problem considers the behavior of a single traveler departing node u for node v, whose impact on prevailing travel times is assumed to be negligible. Thus, the delay functions $t_{ij}^s(x_{ij}^s)$ for each arc state can be replaced by a single travel time t_{ij}^s. Let T_i be a discretization of the set of possible arrival times at node i.

Using the notation defined in Chapter 1, an optimal routing policy $\pi : \Phi \to A$ is sought, where $\pi(i,t,\theta) \in \Gamma(i)$ for all $i \in N$, $t \in T_i$, $\theta \in \Theta_i$. Defining the *node usages* $\eta_i^t(\pi)$ to be the probability that this traveler passes through node i at time t while following policy π, we seek a policy π^* minimizing the expected travel disutility

$$F(\pi) \equiv \sum_{t \in T_v} \eta_v^t(\pi) \tag{3.1}$$

where $f(t)$ represents the disutility of arriving at the destination at time t. Likewise, a vector of *time-dependent arc usages* $x_{ij}^{st}(\pi)$ can be defined, representing the probability that the traveler travels arc (i, j) in state s, starting at time t, when following the policy π. While not difficult, efficiently calculating $\boldsymbol{\eta}$ and $\mathbf{x^t}$ for a given policy is not trivial, and discussed more fully later in Section 3.2.4.

First, however, this approach is placed in the context of past literature in Section 3.2.1, and a fuller discussion of the impact of different disutility functions is given in Section 3.2.2. Section 3.2.3 presents an exact solution algorithm based on label correcting. Following this, an algorithm for calculating the auxiliary variables is discussed.

The section concludes by discussing the possibility of travel routes including cycles in Section 3.2.5, and demonstrating the algorithms in the Seattle network in Section 3.2.6.

3.2.1 Literature Review

Routing under uncertainty has been studied extensively, and problems can be classified according to the nature of the network and the objective being solved. In many applications where arc costs are stochastic, the expected shortest path is often sought. Hall (1986) was the first to investigate the case when arc costs are also time-dependent, and developed a dynamic programming algorithm that allows an adaptive decision to be made at each node, according to the arrival time. A nonpolynomial algorithm developed by Miller-Hooks and Mahmassani (2000) solves the problem of finding a single optimal path exactly, while Fu and Rilett (1998) present a tractable heuristic for the same problem. When adaptive route choice is allowed, Pretolani (2000) provides a more efficient algorithm.

Researchers have also considered incorporating reliability measures into stochastic shortest path algorithms. For instance, Sivakumar and Batta (1994) solve a shortest path problem that constrains the variance of the path cost, while Sen et al. (2001) use a multiobjective approach for normally-distributed arc costs. Robust formulations, such as that in Yu and Yang (1998) or Montemanni and Gambardella (2004) solve a minimax shortest path problem when arc cost distributions are unknown, but upper and lower bounds are available. Fan et al. (2005), on the other hand, find a routing policy that minimizes the probability of arriving at the destination later than a specified arrival time. Another approach involving a desired arrival time is found in Gao (2005), where a weighted sum of expected arrival time before and after the target is minimized.

Other authors develop models allowing nonlinear preferences. Loui (1983) and Eiger et al. (1985) develop procedures for linear and exponential utility functions based on dynamic programming, while Murthy and Sarkar (1996) present an algorithm for decreasing quadratic utility functions. Gabriel and Bernstein (2000) provide a heuristic method for finding "non-additive" shortest paths, and Tsaggouris and Zaroliagis (2004) present such an algorithm for monotone and convex disutility functions.

The importance of including this dimension in route choice is well known. McCord and Villoria (1987) showed that a nonlinear utility specification represented travel behavior better than a linear specification, in a stated-preference experiment. More recently, de Lapparent et al. (2002) studied Parisian peak-hour travel survey data, revealing nonlinearities in user preferences for travel time and cost, along with substantial differences between the morning and evening commutes. Examining a stated preference survey in Austin, Texas, Pinjari and Bhat (2006) also concluded that ignoring these nonlinearities leads to significant losses in modeling fidelity. In the context of mode choice, an intercity travel study conducted by Mandel et al. (1994) yet again concluded that nonlinearities in preferences for level-of-service attributes cannot be ignored.

The information provision in this model closely resembles the "temporal dependency" structure in Waller and Ziliaskopoulos (2002), which presents an algorithm for finding an adaptive routing policy minimizing expected travel cost. Provan (2003) and Gao and Chabini (2006) also present algorithms for more general dependency. This case is more difficult; in fact, Polychronopoulos and Tsitsiklis (1996) and Provan (2003) show that this problem is NP-complete. When arc costs are independent, however, Miller-Hooks (2001) develops a polynomial algorithm for online routing in stochastic, time-dependent networks.

While this dissertation is concerned with nonlinearity in user preferences, the specific source of nonlinearity of interest is uncertainty. Almost all approaches attempt to quantify reliability with a single numerical quantity: for instance, Small et al. (2005) and Liu et al. (2007) use the difference between the 80th- and 50-th percentile travel times, while Pinjari and Bhat (2006) use the maximum additional time that could be needed, compared to an average case.

Amidst this past work in routing in stochastic networks, the contribution of this section is the simultaneous consideration of nonlinear preferences and adaptive routing. Previously, this had only been done for specific disutility functions, as in Gao (2005), or restricted classes of disutility functions, as in Boyles (2006). The approach presented here is fully general and, as special cases, can represent prospect theory, "arriving-on-time" behavior, schedule delay, and many other formulations of behavior under uncertainty.

3.2.2 Disutility Functions

The disutility function f describes the user's preferences, representing risk aversion, schedule delay, or other behavior in the face of uncertain arc delays. By varying the disutility function, different types of user behavior can be described. Possible disutility functions include:

Linear: All increasing linear (or affine) disutility functions are equivalent in the sense that an optimal policy for any such disutility function is also optimal for any other in this class. Thus, all linear disutility functions are equivalent to minimizing $E[t]$, which reduces the problem to the standard online shortest path problem, as described in Waller and Ziliaskopoulos (2002) and Provan (2003). Such a disutility function is applicable when the traveler simply wishes to arrive as quickly as possible, and is risk-neutral regarding uncertain travel times, as might describe weekend shopping trips or other leisure activities where on-time arrival is not a paramount concern.

Deviance: Given a target arrival time t^*, the deviance of a policy π is $E_\pi[(t-t^*)^2]$, that is, $f(t) = (t - t^*)^2$. (Figure 3.1.) This disutility function was proposed in Boyles (2006) as an modification of variance which is suitable for use in online routing algorithms. Since such algorithms require the disutilty function to be completely specified *a priori*, and since variance is defined relative to the mean arrival time, which is endogenous.

Depending on the application, a target arrival time may be clear (as in delivery applications, or in commutes to jobs with a fixed start to the work day.) In other situations where there is no obvious choice for a target arrival time, a reasonable choice of t^* is the expected arrival time of the least-expected time shortest path (linear disutility function) solution $\bar{t}$. In this case, t is not hard to show that the minimum deviance policy necessarily has a lower travel time variance than the least expected-time policy.

Proposition 3.2.1. *If* π^*_{DEV} *is an optimal policy with respect to the*

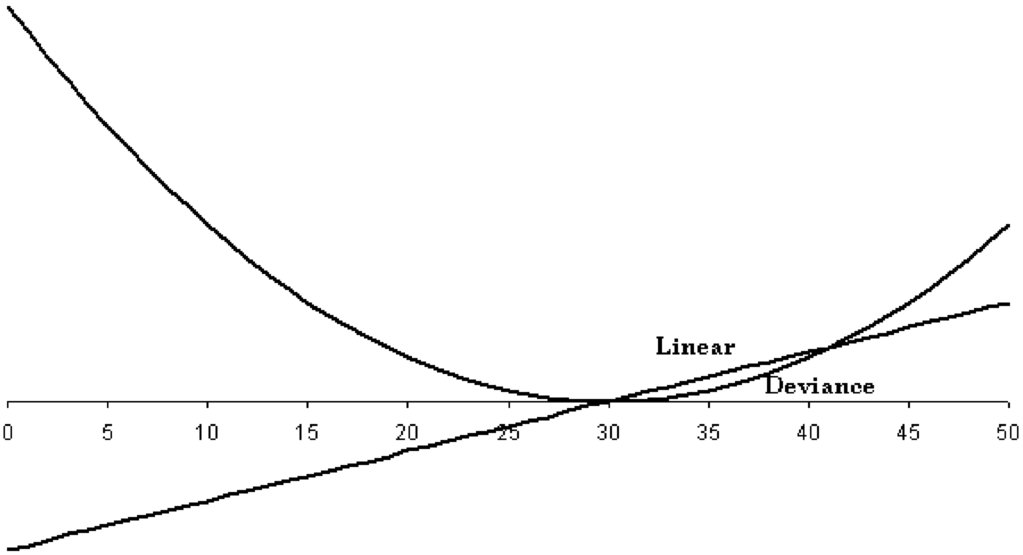


Figure 3.1: Linear disutility function vs. deviance.

disutility function $(t-\bar{t})^2$*, its travel time variance* $\sigma^2(\pi^*_{DEV})$ *is no greater than that of a least expected-time policy* π^*_{LET}*.*

Proof.

$$\begin{aligned}
\sigma^2(\pi^*_{DEV}) &= \sum_{t\in T_v} \eta_v^t(\pi^*_{DEV})t^2 - \left(\sum_{t\in T_v} \eta_v^t(\pi^*_{DEV})t\right)^2 \\
&= \sum_{t\in T_v} \eta_v^t(\pi^*_{DEV})(t-\bar{t})^2 - (\sum_{t\in T_v} \eta_v^t(\pi^*_{DEV})t - \bar{t})^2 \\
&\le \sum_{t\in T_v} \eta_v^t(\pi^*_{DEV})(t-\bar{t})^2 \\
&\le \sum_{t\in T_v} \eta_v^t(\pi^*_{LET})(t-\bar{t})^2 \\
&= \sigma^2(\pi^*_{LET})
\end{aligned}$$

□

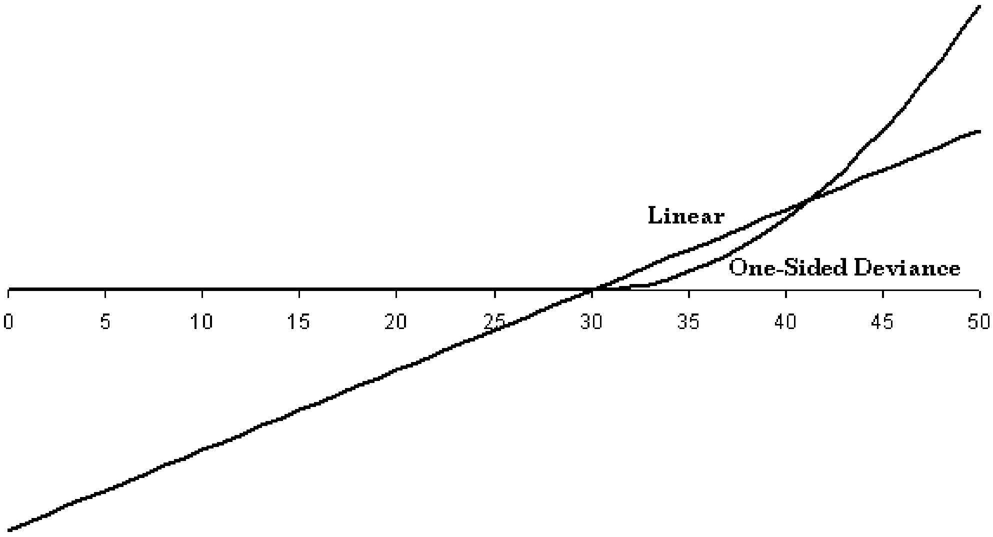


Figure 3.2: Linear disutility function vs. upside deviance.

Deviance is an appropriate disutility function when there is an arrival time that is clearly optimal, and arriving either early or late carries penalties, as in supply chains with high storage costs. Alternately, one may only wish to penalize late arrivals by defining a one-sided deviance $f(t) = ([t - t^*, 0]^+)^2$ which may better describe commuters traveling to work, where late arrivals are penalized in a nonlinear fashion, but no such penalty applies to early arrival. Unlike the two-sided deviance function, no choice of target arrival time can ensure a reduction in variance, compared to the least expected-time policy. One-sided deviance is shown in Figure 3.2

Monotonic Quadratic: A quadratic disutility function increasing over the set of all possible arrival times models a traveler who wishes to arrive as soon as possible, but is either risk-averse regarding uncertain network conditions if this function is convex, or risk-prone if it is concave. This is in contrast to deviance, where arriving at the target time is preferred to earlier arrivals, and to one-sided deviance, where any arrival before the target is equally preferable. However, unlike a linear disutility function, the traveler behaves in a risk-averse manner.

As described in Boyles and Waller (2007a), by exploiting the equivalence of utility functions under affine transformations, it is possible to parameterize monotonic quadratic disutility functions by a single scalar $k \in [-2, 2]$, representing the change in the derivative of f between the earliest and latest possible arrival times t_m and t_M. The disutility function can then be written

$$f(t) = \frac{k}{2}t^2 + \left(1 - \frac{k}{2}\right)\left(t_M - t_m - \frac{2t_m}{t_M - t_m}\right)t \tag{3.2}$$

where $k > 0$ corresponds to risk aversion, $k = 0$ to risk neutrality, and $k < 0$ to risk-prone behavior. Several disutility functions with varying k values are shown in Figure 3.3, with suitable affine transformations applied to each function so they can be viewed on a comparable scale.

Box-Cox: Some researchers (e.g., de Lapparent et al., 2002) have used the transformation $f(t) = (t^\lambda - 1)/\lambda$, developed by Box and Cox (1964) to describe user behavior under uncertainty. λ is a parameter in the transformation; it can be shown that this function approaches $\log t$ as

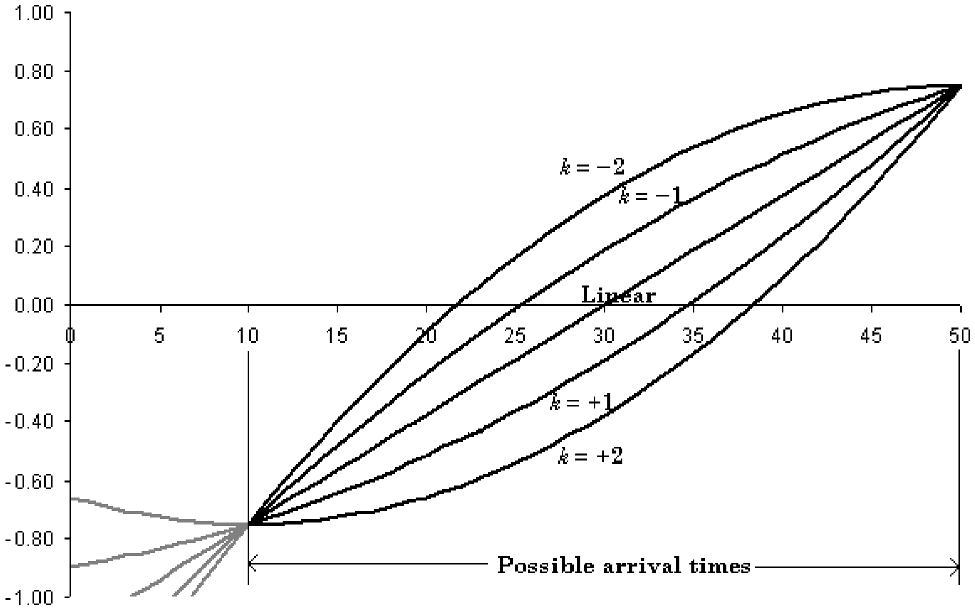


Figure 3.3: Various quadratic disutility functions.

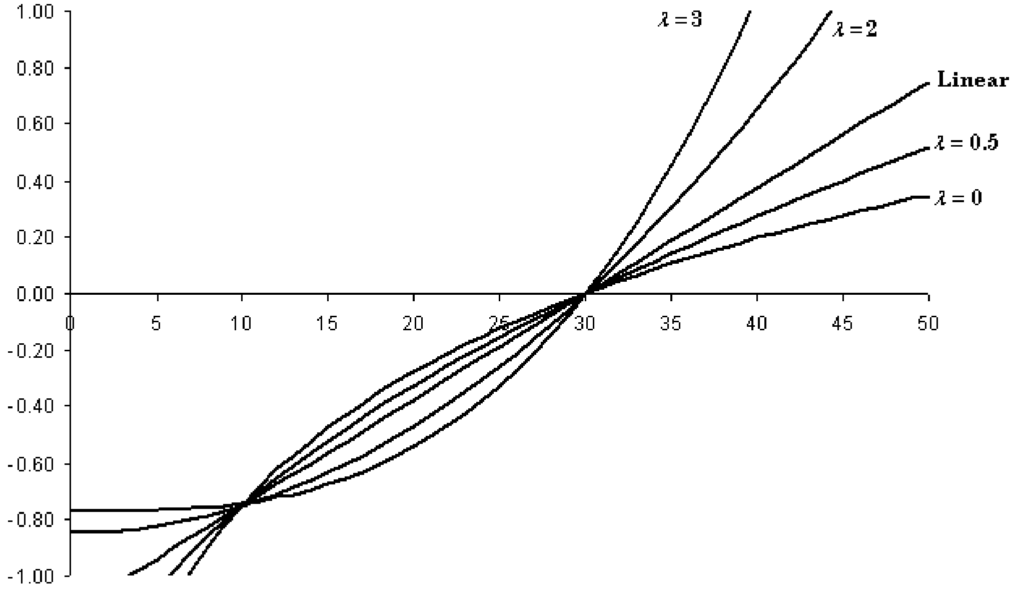


Figure 3.4: Various Box-Cox disutility functions.

$\lambda \to 0$. $\lambda > 1$ corresponds to risk aversion; $\lambda < 1$ to risk-prone behavior. One advantage of this function is that λ can be estimated from revealed preference data, allowing risk preferences to be directly measured, rather than assuming an exact functional form *a priori*. de Lapparent et al. (2002) found values of λ in the range 0.69–1.84 in various specifications of mode choice models applied to travel survey data collected in Paris, France. Various Box-Cox disutility functions are plotted in Figure 3.4.

Arriving On Time: A series of papers (Fan et al., 2005; Fan and Nie, 2006; Nie and Fan, 2006; Nie and Wu, 2009) is addressed at variations of the "arriving on time" problem, where travelers wish to maximize the probability of arriving at the destination no later than a threshold time t^*, after which the traveler is considered late. This behavioral motivation can be represented in the current framework by using the disutility function $f(t) = I(t > t^*)$ where $I(\cdot)$ is an indicator function equal to unity if its argument is true, and zero otherwise. From the standpoint of implementation, it is useful to multiply this indicator function by a constant slightly greater than one:

$$f(t) = (1 + \epsilon)I(t > t^*) \tag{3.3}$$

Otherwise, once travelers are late, they have no incentive to arrive at all, which can cause algorithmic complications.

Unlike the other disutility functions discussed so far, the arriving-on-time function is discontinuous. This does not pose a problem for the routing algorithm; however, for the equilibrium model presented in Section 3.3, continuous disutility functions are needed to ensure existence of a fixed point solution. In such cases, it is not difficult to "smooth" this function so it retains the general character of arriving-on-time while satisfying the necessary regularity conditions. The demonstration of the NL-UER equilibrium algorithm in Section 3.3.5 shows how this may be done.

Prospect Theory: Proposed by Kahneman and Tversky (1979), prospect theory provides yet another approach for modeling behavior under uncertainty through a two-phase process: first, possible outcomes are ranked, and a reference point is chosen. The desirability of each outcome is determined by an S-shaped value function, based on its desirability relative to the reference point. This function is usually chosen to represent loss aversion, where outcomes worse than the reference point are penalized more heavily. Next, an action is chosen maximizing the expected value of the value function. Given a reference utility f^* and a value function $v(\beta, e, f^*)$, this approach can be modeled in our framework by choosing the disutility function to be the negative of the value function for all possible outcomes.

3.2.3 Solution Algorithm

Dynamic programming allows optimal policies to be found efficiently. For the case of piecewise polynomial disutility functions, Boyles (2006) developed a recursive method using the binomial theorem to calculate the moments of remaining travel time for each node state, allowing calculation of expected disutility. A more efficient approach, which applies equally well to disutility functions which are *not* piecewise polynomial, is to store a label at each node state indicating the expected disutility obtained by departing that node state, and following the policy developed thus far.

This allows a policy to be constructed one node-state at a time, starting at the destination (where there is no remaining uncertainty), and working backwards through the network, either adding a new node state to the policy, or updating the policy if a better decision is found. With this in mind, algorithm FINDADAPTIVEPOLICY can be properly presented. Let the label $L(i, t)$ denote the expected disutility achieved if one arrives at node i at time t. The algorithm initializes by setting all such labels to ∞ except at the destination, where the labels are initialized to the disutility value corresponding to the arrival time at each such node state.

There are two main ways in which to determine the node labels. Once is to proceed node-wise, using a scan eligible list to maintain the set of nodes which still need to be examined. This set is initialized to the nodes immediately upstream of the destination. Whenever a node i is "scanned" by the algorithm, the optimal arc for all node-states corresponding to node i are chosen by exhaustively examining each alternative, using the disutility labels stored at the downstream nodes. Once the optimal arc choices are determined, the disutility labels at i can be calculated as well, and the nodes upstream of i added to the scan eligible list. This approach is highly reminiscent of label correcting shortest path algorithms (Ahuja et al., 1993).

An alternative approach is to exploit the acyclic nature of the time-expanded graph, noting that the time indices provide a natural topological ordering. One can then proceed in decreasing order of time, starting with the labels at time T, $T-1$, and so forth. This approach was adopted by Chabini (1999) for solving a deterministic time-dependent shortest path problem in transportation networks.

The decreasing order of time approach was found to outperform the node-wise method in the networks tested, often by nearly an order of magnitude. This is consistent with Miller-Hooks (2001), who found that proceeding time-wise is superior in (relatively sparse) transportation networks, while the node-wise approach is superior in (denser) data networks. For this reason, FINDADAPTIVEPOLICY is presented with a time-wise implementation in this dissertation.

As a standard dynamic programming problem, this algorithm always terminates with the optimal policy π for the given disutility function f. A formal presentation of FindAdaptivePolicy is given as Algorithm 1 on page 72.

Care must be taken if the travel times do not align with the time discretization. For instance, in general, when departing node i at time t on arc (i, j), one may arrive at node j at a time t' which is strictly between two arrival times $t_j^1, t_j^2 \in T_j$. In this case, a linear interpolation is used to estimate the expected disutility associated with this action:

$$L(j, t_j^1) + \frac{t' - t_j^1}{t_j^2 - t_j^1}(L(j, t_j^2) - L(j, t_j^1)) \tag{3.4}$$

This replaces $L(j, t + t_{ij}^s)$ in line 20 of FINDADAPTIVEPOLICY. Because the decreasing order of time approach requires labels to be determined strictly based on future states, if $t_j^1 = t$ (that is, the travel time on (i, j) is less than the unit of the time discretization), the travel time must be rounded up to a full time unit. With an appropriately fine discretization (30–60 seconds) this should not be a problem.

In the worst case, the inner summation over all messages θ can dominate the computation time, since $|\Theta|$ is $O(S^m)$. A matrix reduction procedure described in Waller and Ziliaskopoulos (2002) can reduce this to $O(Sm)$. However, this requires additional overhead, and as transportation networks

Algorithm 1 FindAdaptivePolicy($\mathbf{t}, \mathbf{f}, v$)

```
{Arguments $\mathbf{t}$ and $f$ contain state-dependent travel times $t^s_{ij}$ and the disutility function and specify the destination $v$}
for all $i \in N \backslash v$, $t \in T_i$ do
    $L(i,t) \leftarrow \infty$
end for
for all $t \in T_v$ do
    $L(v,t) \leftarrow f(t)$
end for
$t \leftarrow T - 1$
while $t \geq 0$ do
    for all $i \in N$ do
        for all $\theta \in \Theta_i$ do
            $temp_L \leftarrow 0$
            $temp^\theta \leftarrow \infty$
            for all $j \in \Gamma(i)$ do
                $temp^\theta_j \leftarrow 0$
                for all $s_{ij} \in \theta$ do
                    if $t + t^s_{ij} > T$ then
                        $temp^\theta_j \leftarrow \infty$
                    else
                        $temp^\theta_j \leftarrow temp^\theta_j + L(j, t + t^s_{ij}) p^{s\theta}_{ij}$
                    end if
                end for
                if $temp^\theta_j < temp^\theta$ then
                    $temp^\theta \leftarrow temp^\theta_j$
                    $\pi(i,t,\theta) \leftarrow j$
                    $SEL \leftarrow SEL \cup \Gamma^{-1}(i)$
                end if
            end for
            $temp_L \leftarrow temp_L + temp^\theta_\pi(i,t,\theta) Pr(\theta)$
        end for
        $L(i,t) \leftarrow temp_L$
    end for
    $t \leftarrow t - 1$
end while
return $\pi$
```

are sparse, $|\Theta|$ is rarely large, and bounded regardless of network size. For these reasons, the more straightforward presentation is given in Algorithm 1.

3.2.4 Determining Node and Arc Usages

Calculating the probability that travelers using a given policy will pass a certain node at a certain time, or traverse a certain arc at a certain time, is highly useful. First, it allows any policy to be evaluated according to any disutility function, regardless of its optimality. This way, multiple attributes of policies can be studied (such as the probability that a least-expected time policy is "on time", or the variance of an optimal Box-Cox policy).

Second, this plays a fundamental role in the equilibrium model presented in Section 3.3, where travel times depend on the total arc usage over the time horizon $x_{ij}^s = \sum_t x_{ij}^{st}$. This calculation must be performed many times to find an equilibrium, so an efficient algorithm is needed.

The node usages η_i^t and time-dependent arc usages x_{ij}^{st} can be calculated by another dynamic programming procedure LoadPolicy, this time working in *increasing* order of time. Initially, at time $t = 0$, $\eta_u^0 = 1$ at the origin, and $\eta_i^0 = 0$ everywhere else. For each message θ that can be received at u at time zero, the policy π is consulted, and the probability of observing θ can be added to $x_{\pi(u,0,\theta)}^{s,0}$, proportional to $p_{\pi(u,0,\theta)}^{s\theta}$, to reflect the probability that this arc will be traversed. At the same time, the usage for the downstream node is incremented by the same amount.

The algorithm then proceeds to the next time interval, and considers any node with a positive usage for $t = 1$. The probability of using each adjacent arc is determined from the message structure and the policy, and the appropriate arc usages and downstream node usages are updated accordingly.

As with FINDADAPTIVEPOLICY, a complication arises if the travel times do not align with the time discretization. A "reverse interpolation" is used to determine which node usages are incremented, and by how much. Specifically, if the arrival time t' at node j lies strictly between the time intervals $t_j^1, t_j^2 \in T_j$, the node usage $\eta_j^{t_j^1}$ receives $(t_j^2 - t')/(t_j^2 - t_j^1)$ of the full increment, while $\eta_j^{t_j^2}$ receives $(t' - t_j^1)/(t_j^2 - t_j^1)$. As before, travel times smaller than the unit of the time discretization must be rounded up to one full unit. This requires an adjustment in line 14 of Algorithm 2.

Keeping in mind the future application of LOADPOLICY to an equilibrium problem, an alternate interpretation of this algorithm is useful: instead of representing the probabilities of traversing nodes and arcs at different times, assume that a single, infinitely divisible unit of flow is departing u, and choosing arcs according to the policy. By the independence assumption, this unit of flow will be split according to the probabilities given by the message structure and arc states, exactly as the probabilities for the node and arc usages are calculated.

Thus, LOADPOLICY takes the OD matrix $\mathbf{D}$ as an argument, and simultaneously loads all of the demand destined for node v using policy π. In this case, the node and arc usages represent the aggregate uses by all such travelers. The node and arc usages for a single policy can easily be obtained by setting all elements of $\mathbf{D}$ to zero, exept for the uv-th, which is set to unity.

Finally, note that an incidence matrix A can be constructed, associating each policy $\pi \in \cup_{(u,v)\in D}\Pi_{uv}$ with the arc usages $x_{ij}^{s} = \sum_t x_{ij}^{st}$ associated with unit demand from u to v. Each element $a_{\pi,ijs}$ denotes the proportion of travelers using policy π who will use arc (i,j) in state s. This matrix is far too large to construct and use explicitly; however, it provides a useful notational shorthand, and portions of it can be constructed as necessary.

Algorithm 2 LoadPolicy$(\pi, v, \mathbf{D})$

```
{Arguments: policy $\pi$, destination $v$, OD demand $\mathbf{D}$)}
$\mathbf{x^t} \leftarrow \mathbf{0} \qquad \forall t \in \{0, 1, \ldots, T\}$
$\boldsymbol{\eta} \leftarrow \mathbf{0}$
for all $(u, v) \in D$ do
    $\eta_i^0 \leftarrow d_{uv}$
end for
$t \leftarrow 0$
while $t < T$ do
    for all $i \in N : \eta_i^t > 0$ do
        for all $\theta \in \Theta_i$ do
            $(i, j) \leftarrow \pi(i, t, \theta)$
            for all $s \in S_{ij}$ do
                $x_{ij}^{st} \leftarrow x_{ij}^{st} + \rho_i^\theta p_{ij}^{s\theta} \eta_i^t$
                $\eta_j^{t+t_{ij}^s} \leftarrow \eta_j^{t+t_{ij}^s} + \rho_i^\theta p_{ij}^{s\theta} \eta_i^t$
            end for
        end for
    end for
    $t \leftarrow t + 1$
end while
return $(\mathbf{x}, \boldsymbol{\eta})$
```

3.2.5 Cycles and *Contretemps*

Consider the network in Figure 3.5, found in Waller and Ziliaskopoulos (2002), where all arcs have deterministic cost except that leading to the destination 4. If node 2 is an information node, clearly the optimal strategy is to travel to the destination if arc $(2,4)$ has low cost, and to traverse the cycle $(2,3,1,2)$ if arc $(2,4)$ has high cost. In the latter case, the traveler is relying on the independence ("reset") assumption, hoping that arc $(2,4)$ will be in the low cost upon a subsequent traversal. This does not sit well with intuition regarding driver behavior, although in very specific cases it is in fact frequently observed — consider drivers who circle endlessly seeking the perfect parking spot in front of their destination, hoping that a space will be open the next time they pass by. Implicitly, for these drivers, the expected disutility of traversing a cycle is less than that of a longer, deterministic path to the destination (such as parking further away, and walking), which is exactly the condition needed for cycling to occur.

We term this phenomenon of returning to a node previously visited a *contretemps* both because of its counterintuitive nature and because the traveler's action resembles a movement "against time" that the French etymology implies. For some disutility functions, *contretemps* may occur as a matter of routine: if travelers' behavior is defined by schedule delay or deviance, in which early arrival is penalized, rational travelers who are ahead of schedule may opt to "drive around the block another time" so as to arrive closer to the desired time. *Contretemps* are more surprising when disutility functions are

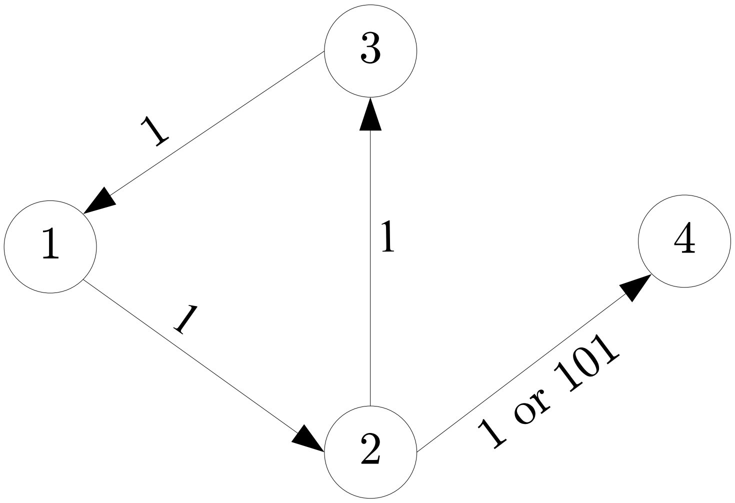


Figure 3.5: Illustration of a *contretemps*

strictly increasing; in this case, *contretemps* can occur only if G is *non-FIFO* in the sense that there exist one or more arcs where the traveler can arrive at the downstream node earlier by departing the upstream node later, with nonzero probability.

Contretemps are more insidious than a first glance might suggest: in the example in Figure 3.5, there is no guarantee that $(2,4)$ will have low cost even after traversing the cycle $(2,3,1,2)$, in which case the cycle is traversed again, and so forth *ad infinitum*. This is problematic because the presence of any finite time horizon T, no matter how large, introduces error by eliminating the rare possibilities of a very large number of cycles being performed. However, the probability of a large number of cycles shrinks relatively quickly (geometrically), as shown in Waller and Ziliaskopoulos (2002), and these authors provide a bound on the error introduced by terminating a node-wise algorithm after a specified number of iterations.

The possibility of *contretemps* implies that the time horizon T imposed on the routing problem may lead to modeling difficulties, since the time horizon is an artifact whose alteration should not choose the optimal policies. Nevertheless, we can form bounds on the difference in expected disutility between the optimal T-constrained policy π^*_T and the optimal unconstrained policy π^*_∞, under the following nonrestrictive assumption:

Assumption 5. *There exists a time* t_0 *such that* $f(t)$ *is increasing for* $t > t_0$*.*

This derivation contrasts with the node-wise algorithmic bounds of Waller and Ziliaskopoulos (2002), providing bounds instead for time-wise online algorithms. Furthermore, the approach is general and can account for nonlinear disutility functions.

Let $\overline{L}$ represent the greatest minimax distance between v and any other node, added to the greatest possible cost which can be experienced on an arc; that is,

$$\overline{L} = \max_{i \in N} \left\{ \min_{p \in P_{iv}} \left\{ \sum_{(i,j) \in p} \max_{s \in S_{ij}} \left\{ c^s_{ij} \right\} \right\} \right\} + \max_{(i,j) \in A} \left\{ \max_{s \in S_{ij}} \left\{ c^s_{ij} \right\} \right\} \tag{3.5}$$

where P_{iv} is the set of simple paths connecting nodes i and v. This quantity is significant, because every pair of nodes is connected by a path whose travel time is no greater than $\overline{L}$ with probability 1. Further define the sets of arrival times $T_v^1(\tau)$, T_v^2, T_v^3, and T_v^4 as follows:

$$T_v^1(\tau) = T_v \cap [\tau, \infty) \tag{3.6}$$

$$T_v^2 = T_v \cap [t_0 - \overline{L}, \infty) \tag{3.7}$$

$$T_v^3 = T_v \cap [0, t_0 - \overline{L}) \tag{3.8}$$

$$T_v^4 = T_v \cap [t_0 - \overline{L}, t_0] \tag{3.9}$$

We first bound the probability that traveler arrives later than a specified time $\tau > t_0 + \overline{L}$ when following the optimal unconstrained policy π_∞^*.

Lemma 3.2.1. *If $\tau > t_0 + \overline{L}$, the probability that the traveler's trip is completed after time τ when following π_∞^* is no greater than $f(t_0 + \overline{L})/f(\tau)$*

Proof. Consider the following policy: travel deterministically from u to any node which is part of a cycle, then traverse this cycle until the current time exceeds t_0, after which a deterministic minimax path is taken to the destination. Clearly this policy results in arrival at the destination between times t_0 and $t_0 + \overline{L}$, with an expected disutility no greater than $f(t_0 + \overline{L})$. Thus, $E[\pi_\infty^*] \leq f(t_0 + \overline{L})$ as well, and we have

$$\sum_{t \in T_v} \eta_v^t(\pi_\infty^*) f(t) \leq f(t_0 + \overline{L}) \tag{3.10}$$

$$\Rightarrow \sum_{t \in T_v^1(\tau)} \eta_v^t(\pi_\infty^*) f(t) \leq f(t_0 + \overline{L}) \tag{3.11}$$

$$\Rightarrow f(\tau) \sum_{t \in T_v^1(\tau)} \eta_v^t(\pi_\infty^*) \leq f(t_0 + \overline{L}) \tag{3.12}$$

$$\Rightarrow \sum_{t \in T_v^1(\tau)} \eta_v^t(\pi_\infty^*) \leq \frac{f(t_0 + \overline{L})}{f(\tau)} \tag{3.13}$$

$\square$

We can now bound the increase in disutility caused by forcing a trip to end before the time horizon T, assuming $T > t_0$. This is accomplished by constructing a (possibly suboptimal) policy π_T which terminates no later than time T with probability 1. For node-states ϕ whose time component is no greater than $T - \overline{L}$, we choose $\pi_T(\phi) = \pi^*_\infty(\phi)$. When the time component exceeds $T - \overline{L}$, the traveler follows a deterministic minimax path from their current node to the destination. That is, the optimal unconstrained policy is used until time $T - \overline{L}$, after which point a deterministic path is followed.

Theorem 3.2.1.

$$F(\pi^*_T) - F(\pi^*_\infty) \leq f(t_0 + \overline{L}) \left[\frac{f(T)}{f(T - \overline{L})} - 1 \right] \tag{3.14}$$

Proof. First, note that

$$F(\pi^*_\infty) = \sum_{t \in T^2_v} \eta^t_v(\pi^*_\infty) f(t) + \sum_{t \in T^3_v} \eta^t_v(\pi^*_\infty) f(t) \tag{3.15}$$

and

$$F(\pi_T) = \sum_{t \in T^2_v} \eta^t_v(\pi_T) f(t) + \sum_{t \in T^3_v} \eta^t_v(\pi_T) f(t) \tag{3.16}$$

By construction, $\eta^t_v(\pi^*_\infty) = \eta^t_v(\pi_T)$ for $t \in T^3_u$ and thus

$$\sum_{t \in T^2_v} \eta^t_v(\pi^*_\infty) f(t) = \sum_{t \in T^2_v} \eta^t_v(\pi_T) f(t) \tag{3.17}$$

Since f is increasing on $[t_0, T]$, we have

$$F(\pi_T) \leq f(T) \sum_{t \in T^2_v} \eta^t_v(\pi^*_\infty) f(t) + \sum_{t \in T^3_v} \eta^t_v(\pi^*_\infty) f(t). \tag{3.18}$$

Furthermore, $F(\pi_T^*) \leq F(\pi_T)$ implies

$$
\begin{aligned}
F(\pi_T^*) - F(\pi_\infty^*) &\leq F(\pi_T) - F(\pi_\infty^*) && (3.19)\\
&\leq \sum_{t \in T_v^2} \eta_v^t(\pi_\infty^*)[f(T) - f(t)] && (3.20)\\
&\leq \sum_{t \in T_v^4} \eta_v^t(\pi_\infty^*)[f(T) - f(t)] && (3.21)
\end{aligned}
$$

with the last inequality true because $f(T) - f(t) < 0$ for $t \in T_u^2 \backslash T_u^4$. Continuing, we have

$$
\begin{aligned}
F(\pi_T^*) - F(\pi_\infty^*) & \sum_{t \in T_v^4} \eta_v^t(\pi_\infty^*)[f(T) - f(T - \overline{L})] && (3.22)\\
&\leq \frac{f(t_0 + \overline{L})}{f(T - \overline{L})}[f(T) - f(T - \overline{L})] && (3.23)\\
&= f(t_0 + \overline{L}) \left[\frac{f(T)}{f(T - \overline{L})} - 1 \right] && (3.24)
\end{aligned}
$$

where the last inequality follows from Lemma 3.2.1. □

The asymptotic properties of this bound are of interest. Particularly, if

$$
\lim_{T \to \infty} \frac{f(T)}{f(T - \overline{L})} = 1 \qquad (3.25)
$$

then any error introduced by the time horizon can be made arbitrarily small by extending the analysis period. Unfortunately, this is not always the case. If $f(t) = e^T$, for instance, $f(T)/f(T - \overline{L}) = e^{\overline{L}} > 1$ and the bound does not shrink with T.

However, the bound does in fact vanish asymptotically for several important cases.

Corollary 3.2.1. (Concave disutility.) *If $f(t)$ is concave and differentiable for sufficiently large t, the limit (3.25) is satisfied.*

Proof.

$$1 \le \frac{f(T)}{f(T-\overline{L})} \le \frac{f(T-\overline{L}) + \overline{L}f'(T-\overline{L})}{f(T-\overline{L})} = 1 + \overline{L}\frac{f'(T-\overline{L})}{f(T-\overline{L})} \tag{3.26}$$

As $f'(T-\overline{L})$ is decreasing and nonnegative, and as $f(T-\overline{L}) \to \infty$ as $T \to \infty$, the result immediately follows. □

Corollary 3.2.2. (Polynomial disutility.) *If $f(t)$ is a polynomial for sufficiently large t, the limit (3.25) is satisfied.*

Proof. A routine application of l'Hospital's Rule establishes the result. □

Corollary 3.2.3. (Box-Cox.) *If $f(t) = (t^\lambda - 1)/\lambda$ for some $\lambda \in \mathbb{R}^+$, the limit (3.25) is satisfied.*

Proof. Trivial. □

A more general condition can be given, using bounds on f and its first derivative.

Corollary 3.2.4. *Assume that $f(t)$ is differentiable for large t, and that there exist constants $C_1, C_2 \in \mathbb{R}$, $B_1, B_2 \in \mathbb{R}^+$ such that $f(t) \ge B_1 t^{C_1}$ and $f'(t) \le B_2 t^{C_2}$ for sufficiently large t. If $C_1 > C_2$, the limit (3.25) is satisfied.*

Proof.

$$1 \leq \frac{f(T)}{f(T-\overline{L})} = 1 + \overline{L}\frac{f'(T-\lambda\overline{L})}{f(T-\overline{L})} \tag{3.27}$$

for some $\lambda \in [0,1]$ by the mean value theorem. From the conditions of the theorem, this implies

$$0 \leq \frac{f(T)}{f(T-\overline{L})} - 1 \leq \overline{L}\frac{B_1 T^{C_1}}{B_2(T-\overline{L})^{C_2}} \tag{3.28}$$

The rightmost quantity vanishes in the limit by l'Hospital's rule, proving the corollary. □

To show how this bound can be used in practice, consider the linear disutility function $f(t) = t$. This function is always increasing, so $t_0 = 0$, and, by substitution into (3.14), the error introduced by forcing trips to end by time T is no greater than

$$\overline{L}\left[\frac{T}{T-\overline{L}} - 1\right] = \frac{\overline{L}^2}{T-\overline{L}} \tag{3.29}$$

If one is interested in a time horizon long enough to reduce this error below some threshold ϵ, one can set ϵ equal to this bound and solve for T, yielding

$$T \geq \overline{L} + \frac{\overline{L}}{\epsilon} \tag{3.30}$$

as a sufficient condition.

Similarly, for the deviance disutility function $f(t) = (t - t^*)^2$, f is increasing for $t \geq t_0 = t^*$. Again substituting into (3.14) and performing some algebraic manipulations, the error ϵ is seen to satisfy

$$\epsilon \leq \frac{\overline{L}^4 + 2\overline{L}^3(T-t^*)}{(T-\overline{L}-t^*)^2} \tag{3.31}$$

or, inverting this equation, a time horizon satisfying

$$T \geq t^* + \overline{L} + \frac{\overline{L}^2}{\epsilon}\left(\overline{L} + \sqrt{\overline{L} + 3\epsilon}\right) \tag{3.32}$$

is sufficient to reduce the error below ϵ.

In general, it is not possible to invert the bound (3.14) analytically, although the minimum T corresponding to a specific ϵ can usually be found numerically without much difficulty using a line search algorithm (see, for instance, Chapra and Canale, 2002).

3.2.6 A Small Demonstration

Referring to the simplified Seattle network of Figure 2.6, consider a traveler attempting to drive from Everett to Tacoma, desiring to arrive $t^* = 70$ minutes after departure. With this behavior assumption, the deviance disutility function is appropriate: $f(t) = (t - t^*)^2$ Arcs can exist in one of two states, no incident (NI) or incident present (IP); the probabilities of these state occurences, as well as the corresponding travel times, can be found in Table 3.1. For the sake of illustration, a 120-minute time horizon is adopted, with a 5-minute discretization. Note that this forces all states with a travel time less than 5 minutes to be rounded up to ensure progression from one time interval to the next — in practice, the input data would warrant a finer discretization, but larger time intervals are more amenable to basic understanding, which is the intent of this section.

The policy is then determined in decreasing order of time, along with the corresponding labels $L(\cdot, \cdot)$; the reader may find it helpful to refer to Tables 3.2 and 3.3 when following along. Initially, the labels are initialized to $(t - t^*)^2$ at the destination, and to ∞ everywhere else. Next, the labels corresponding to the penultimate arrival time $t = 115$ are considered; since it is impossible to depart any node at this time and ensure arrival in Tacoma strictly before the time horizon, these labels are set to ∞ as well. Proceeding to $t = 110$, we see that Tacoma is reachable from nodes 9 and 10 via the arcs $(9, T)$ and $(10, T)$. Although the travel time on these arcs is only one minute, the time discretization inflates these to five minutes, resulting in arrival at the destination at time $t = 115$ and leading to a disutility of $(115 - 70)^2 = 2025$ units.

The next few time intervals are similar; a more interesting event happens at $t = 90$ when scanning node 8. This node is adjacent to three arcs: $(8, 6)$, $(8, 7)$, and $(8, 10)$. Nodes 6 and 10 still have infinite disutility at all future arrival times; but choosing arc $(8, 10)$ leads to arrival at node 10 at time $t = 100$ with probability 0.99, and at time 107 with probability 0.01. Interpolating the labels for node 10, this results in expected disutility of 1225 with probability 0.99, and of 1770 with probability 0.1, and thus $L(8, 90) = 1225 \times 0.99 + 1770 \times 0.01 = 1230$ and $\pi(8, 90, \cdot) = (8, 10)$.

Continuing further, the algorithm eventually scans node 7 at time $t = 85$. This node is adjacent to the arcs $(7, 5)$, $(7, 8)$, and $(7, 9)$; since node 5 still has infinite disutility, the first of these three arcs is ignored. Then, there are four possible messages that can be received at this node:

1. There is no incident on either $(7, 8)$ or $(7, 9)$.
2. There is an incident on $(7, 8)$, but not on $(7, 9)$.
3. There is an incident on $(7, 9)$, but not on $(7, 8)$.
4. There are incidents on both $(7, 8)$ and $(7, 9)$.

These four messages are observed with probabilities 0.76, 0.04, 0.19, and 0.01, respectively.

Consider each of these messages in turn. In the first case, choosing arc $(7, 9)$ implies arrival at node 9 at time 98, with (interpolated) expected disutility 1095, while choosing arc $(7, 8)$ implies arrival at node 8 at time 90 (rounding the travel time up to 5 minutes), with expected disutility 1230. Thus, for this message, the best choice is arc $(7, 9)$. For the second message, choosing arcs $(7, 8)$ and $(7, 9)$ yield expected disutilities of 1095 and 1230, respectively, and again $(7, 9)$ is the superior choice. For the third message, arc $(7, 8)$ is best, with expected disutility 1230; the same is true for the fourth message. The policy is set accordingly, and the label $L(7, 85)$ is set to

$$L(7, 85) = 1095 \times (0.76 + 0.04) + 1230 \times (0.19 + 0.01) = 1122 \qquad (3.33)$$

The algorithm proceeds similarly until all nodes have been processed at $t = 0$, terminating with the optimal policy and expected disutilities for all nodes and departure times.

Table 3.1: Data for NL-OSP demonstration on Seattle network

(i,j)	t_{ij}^{NI}	t_{ij}^{IP}	p_{ij}^{NI}	p_{ij}^{IP}
(1,2)	7	12	0.82	0.18
(2,1)	7	12	0.82	0.18
(2,3)	15	26	0.82	0.18
(2,4)	18	35	0.92	0.08
(3,2)	14	24	0.82	0.18
(3,4)	11	20	0.75	0.25
(3,5)	4	7	0.87	0.13
(4,2)	17	28	0.92	0.08
(4,3)	11	19	0.75	0.25
(4,6)	3	7	0.91	0.09
(5,3)	4	7	0.91	0.09
(5,6)	11	21	0.99	0.01
(5,7)	15	24	0.87	0.13
(6,4)	3	5	0.91	0.09
(6,5)	11	19	0.85	0.15
(6,8)	10	15	0.95	0.05
(7,5)	13	21	0.91	0.09
(7,8)	2	3	0.95	0.05
(7,9)	13	25	0.8	0.2
(8,6)	9	14	0.95	0.05
(8,7)	2	3	0.95	0.05
(8,10)	10	17	0.99	0.01
(9,7)	13	23	0.8	0.2
(10,8)	11	15	0.99	0.01
(B,6)	5	5	1	0
(E,1)	5	5	1	0
(S,5)	5	5	1	0
(9,T)	5	5	1	0
(10,T)	5	5	1	0

Table 3.2: Final node labels for NL-OSP in Seattle

$\downarrow t \| n \rightarrow$	B	E	S	T	1	2	3	4	5	6	7	8	9	10
0	0.44	2.65	0.44	4900	1.17	0.5	0.44	0.44	0.45	0.44	0.44	0.45	2.42	0.54
5	0.45	7.01	0.57	4225	2.65	0.61	0.46	0.44	0.44	0.44	0.57	0.44	1.76	0.65
10	0.44	24.18	0.46	3600	7.01	1.45	0.48	0.45	0.57	0.45	0.44	0.57	8.37	1.26
15	0.61	∞	1.53	3025	24.18	3.24	0.58	0.45	0.46	0.44	1.47	0.44	5.73	3.01
20	0.45	∞	0.58	2500	∞	7.28	0.73	0.5	1.53	0.61	0.44	1.47	10.48	10.2
25	1.77	∞	13.6	2025	∞	31.2	1.52	0.61	0.58	0.45	12.78	0.44	7.15	3.92
30	0.61	∞	1.52	1600	∞	97.3	2.49	1.18	13.48	1.77	0.44	12.3	10.5	13.7
35	16.3	∞	21.1	1225	∞	∞	24.4	1.83	1.52	0.61	17.18	0.44	16.3	4.01
40	1.8	∞	84.1	900	∞	∞	90.0	2.85	21.1	16.4	0.44	17.18	36.8	13.72
45	30.0	∞	197	625	∞	∞	206	32.8	84.1	1.84	17.2	0.55	98.4	5.9
50	108	∞	360	400	∞	∞	371	113	197	30.0	0.55	17.2	159	42
55	236	∞	573	225	∞	∞	∞	243	360	108	17.3	0.55	100	100
60	414	∞	∞	100	∞	∞	∞	423	573	236	76.4	26.3	25	25
65	642	∞	∞	25	∞	∞	∞	653	∞	414	186	102	0	0
70	921	∞	∞	0	∞	∞	∞	∞	∞	643	345	228	25	25
75	∞	∞	∞	25	∞	∞	∞	∞	∞	921	554	403	100	100
80	∞	∞	∞	100	∞	∞	∞	∞	∞	∞	813	629	225	225
85	∞	∞	∞	225	∞	∞	∞	∞	∞	∞	1122	905	400	400
90	∞	∞	∞	400	∞	∞	∞	∞	∞	∞	∞	1230	625	625
95	∞	∞	∞	625	∞	∞	∞	∞	∞	∞	∞	∞	900	900
100	∞	∞	∞	900	∞	∞	∞	∞	∞	∞	∞	∞	1225	1225
105	∞	∞	∞	1225	∞	∞	∞	∞	∞	∞	∞	∞	1600	1600
110	∞	∞	∞	1600	∞	∞	∞	∞	∞	∞	∞	∞	2025	2025
115	∞	∞	∞	2025	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
120	∞	∞	∞	2500	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞

Table 3.3: Final policy for NL-OSP in Seattle (selected nodes; $t \leq 35$)

i	Incidents	0	5	10	15	20	25	30	35
E		(E,1)	(E,1)	(E,1)	(E,1)	(E,1)	(E,1)	(E,1)	(E,1)
1		(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)
1	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)
2		(2,4)	(2,3)	(2,4)	(2,3)	(2,4)	(2,4)	(2,4)	(2,4)
2	(2,1)	(2,4)	(2,3)	(2,4)	(2,3)	(2,4)	(2,4)	(2,4)	(2,4)
2	(2,3)	(2,4)	(2,3)	(2,4)	(2,4)	(2,4)	(2,4)	(2,4)	(2,4)
2	(2,1) (2,3)	(2,4)	(2,3)	(2,4)	(2,4)	(2,4)	(2,4)	(2,4)	(2,4)
2	(2,4)	(2,1)	(2,3)	(2,3)	(2,3)	(2,3)	(2,3)	(2,3)	(2,3)
2	(2,1) (2,4)	(2,3)	(2,3)	(2,3)	(2,3)	(2,3)	(2,3)	(2,3)	(2,3)
2	(2,3) (2,4)	(2,1)	(2,4)	(2,4)	(2,4)	(2,3)	(2,4)	(2,4)	(2,4)
2	(2,1) (2,3) (2,4)	(2,3)	(2,4)	(2,4)	(2,4)	(2,3)	(2,4)	(2,4)	(2,4)
4		(4,6)	(4,3)	(4,6)	(4,3)	(4,6)	(4,6)	(4,6)	(4,6)
4	(4,2)	(4,6)	(4,3)	(4,6)	(4,3)	(4,6)	(4,6)	(4,6)	(4,6)
4	(4,3)	(4,6)	(4,6)	(4,6)	(4,6)	(4,6)	(4,6)	(4,6)	(4,6)
4	(4,2) (4,3)	(4,6)	(4,6)	(4,6)	(4,6)	(4,6)	(4,6)	(4,6)	(4,6)
4	(4,6)	(4,3)	(4,6)	(4,3)	(4,6)	(4,6)	(4,6)	(4,6)	(4,6)
4	(4,2) (4,6)	(4,3)	(4,6)	(4,3)	(4,6)	(4,6)	(4,6)	(4,6)	(4,6)
4	(4,3) (4,6)	(4,3)	(4,6)	(4,3)	(4,6)	(4,6)	(4,6)	(4,6)	(4,6)
4	(4,2) (4,3) (4,6)	(4,3)	(4,6)	(4,3)	(4,6)	(4,6)	(4,6)	(4,6)	(4,6)
6		(6,5)	(6,8)	(6,5)	(6,8)	(6,5)	(6,8)	(6,4)	(6,8)
6	(6,4)	(6,5)	(6,8)	(6,5)	(6,8)	(6,5)	(6,8)	(6,4)	(6,8)
6	(6,5)	(6,8)	(6,8)	(6,4)	(6,8)	(6,8)	(6,8)	(6,4)	(6,8)
6	(6,4) (6,5)	(6,8)	(6,8)	(6,4)	(6,8)	(6,8)	(6,8)	(6,4)	(6,8)
6	(6,8)	(6,8)	(6,4)	(6,8)	(6,4)	(6,8)	(6,4)	(6,8)	(6,4)
6	(6,4) (6,8)	(6,8)	(6,4)	(6,8)	(6,4)	(6,8)	(6,4)	(6,8)	(6,4)
6	(6,5) (6,8)	(6,8)	(6,4)	(6,8)	(6,4)	(6,8)	(6,4)	(6,8)	(6,4)
6	(6,4) (6,5) (6,8)	(6,8)	(6,4)	(6,8)	(6,4)	(6,8)	(6,4)	(6,8)	(6,4)
8		(8,7)	(8,7)	(8,7)	(8,7)	(8,6)	(8,7)	(8,6)	(8,7)
8	(8,6)	(8,6)	(8,7)	(8,6)	(8,7)	(8,6)	(8,7)	(8,6)	(8,7)
8	(8,7)	(8,7)	(8,7)	(8,7)	(8,7)	(8,6)	(8,7)	(8,6)	(8,7)
8	(8,6) (8,7)	(8,6)	(8,7)	(8,6)	(8,7)	(8,6)	(8,7)	(8,6)	(8,7)
8	(8,10)	(8,10)	(8,7)	(8,10)	(8,7)	(8,10)	(8,7)	(8,6)	(8,7)
8	(8,6) (8,10)	(8,6)	(8,7)	(8,6)	(8,7)	(8,6)	(8,7)	(8,6)	(8,7)
8	(8,7) (8,10)	(8,10)	(8,7)	(8,10)	(8,7)	(8,10)	(8,7)	(8,6)	(8,7)
8	(8,6) (8,7) (8,10)	(8,6)	(8,7)	(8,6)	(8,7)	(8,6)	(8,7)	(8,6)	(8,7)
9		(9,7)	(9,7)	(9,7)	(9,7)	(9,7)	(9,7)	(9,7)	(9,7)
9	(9,7)	(9,7)	(9,7)	(9,7)	(9,7)	(9,7)	(9,7)	(9,7)	(9,7)
10		(10,8)	(10,8)	(10,8)	(10,8)	(10,8)	(10,8)	(10,8)	(10,8)
10	(10,8)	(10,8)	(10,8)	(10,8)	(10,8)	(10,8)	(10,8)	(10,8)	(10,8)

Table 3.4: Node arrivals for NL-OSP in Seattle

$\downarrow t \mid n \rightarrow$	B	E	S	T	1	2	3	4	5	6	7	8	9	10
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0.49	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0.44	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0.07	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0.03	0.18	0	0	0	0	0	0
30	0	0	0	0	0	0	0.36	0.3	0.01	0.17	0	0	0	0
35	0	0	0	0	0	0	0	0.25	0.32	0.3	0	0	0	0
40	0	0	0	0	0	0	0	0.09	0.01	0.26	0	0	0	0
45	0	0	0	0	0	0	0	0.01	0	0.13	0	0.29	0	0
50	0	0	0	0	0	0	0	0.01	0	0.03	0.57	0.25	0	0
55	0	0	0	0	0	0	0	0	0	0.01	0.25	0.7	0	0
60	0	0	0	0	0	0	0	0	0	0	0	0.08	0	0
65	0	0	0	0	0	0	0	0	0	0	0	0.01	0.08	0.7
70	0	0	0	0.78	0	0	0	0	0	0	0	0	0.12	0.09
75	0	0	0	0.21	0	0	0	0	0	0	0	0	0	0.01
80	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0
85	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0
95	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
105	0	0	0	0	0	0	0	0	0	0	0	0	0	0
110	0	0	0	0	0	0	0	0	0	0	0	0	0	0
115	0	0	0	0	0	0	0	0	0	0	0	0	0	0
120	0	0	0	0	0	0	0	0	0	0	0	0	0	0

To obtain the arc usages associated with this policy, apply LoadPolicy. This algorithm maintains two sets of auxiliary variables: the node usages η_i^t, representing the probability that node i will be visited at time t; and the time-dependent arc-state usages x_{ij}^{st}, representing the probability that arc (i, j) will be traversed in state s, departing node i at time t. Since the delay functions only depend on the total arc-state usages, the time-dependent usages x_{ij}^{st} ultimately map to the arc-state usages x_{ij}^s by summation:

$$x_{ij}^s = \sum_{t \in T_i} x_{ij}^{st} \tag{3.34}$$

The algorithm initializes by setting both sets of auxiliary variables to zero everywhere, with the exception of $\eta_E^0 = 1$ to mark the trip's starting location and time. Arc $(E, 1)$ must be followed to node 1, resulting in $x_{E1}^{NI,0} = 1$ and $\eta_1^5 = 1$. From here, arc $(1, 2)$ must be followed. There is an incident with probability 0.18, and no incident with probability 0.82; hence, $x_{12}^{NI,5} = 0.82$ and $x_{12}^{IP,5} = 0.18$. In the former case, the flow arrives at node 2 at time 12; in the latter case, at time 17. Performing a reverse interpolation and adding, this sets $\eta_2^{10} \leftarrow 0.82 \times 3/5 = 0.49$, $\eta_2^{15} \leftarrow 0.82 \times 2/5 + 0.18 \times 3/5 = 0.44$, and $\eta_2^{20} \leftarrow 0.18 \times 2/5 = 0.07$.

Upon arrival at node 2, one of eight possible messages will be received; for each of these, the policy is consulted to see which arc is taken, and the appropriate amount of flow is added to the appropriate time-dependent arc usage and node usage. To this point, each node is tracked individually; in general, all of the nodes corresponding to each arrival time are scanned consecutively, with the algorithm proceeding in increasing order of arrival time. The remaining node usages and time-dependent arc usages are shown in Tables 3.4 and 3.5.

3.3 Equilibrium

The equilibrium problem builds on the routing problem by considering the collective behavior of self-interested travelers (where the definition of "self-interested" varies by user class). Unlike the routing problem, the travel times must be internally determined by the demand for travel on a particular arc.

The problem considered in this dissertation can be defined by analogy to the classical static, deterministic user equilibrium problem, where one seeks an assignment of travelers to paths which satisfies demand, and where each used path has equal and minimal cost among paths available to that origin-destination pair. Instead of an assignment of travelers to paths, here we seek an assignment of travelers to *policies* which satisfies demand, and where each used policy has equal and minimum expected disutility among policies available to that origin-destination pair and user class. This problem is termed nonlinear user equilibrium with recourse (NL-UER), and the key decision variables are the number of travelers from each OD pair (u, v) and user class q using the policy $\pi \in \Pi_{uv}$; we write this as y_q^π, where the OD pair is implicitly specified by π.

Formally, the NL-UER problem is defined as follows. First, the allowable policy flow vectors are defined.

Table 3.5: Final arc usages for NL-OSP in Seattle (selected arcs)

(i,j)	State	0	5	10	15	20	25	...	120	x_{ij}^s
(E,1)	NI	1	0	0	0	0	0	...	0	1
(1,2)	NI	0	0.82	0	0	0	0	...	0	0.82
(1,2)	IP	0	0.18	0	0	0	0	...	0	0.18
(2,3)	NI	0	0	0.03	0.36	0	0	...	0	0.39
(2,3)	IP	0	0	0	0	0	0	...	0	0
(2,4)	NI	0	0	0.45	0.07	0.07	0	...	0	0.59
(2,4)	IP	0	0	0.01	0.01	0	0	...	0	0.01
(3,4)	NI	0	0	0	0	0	0.02	...	0	0.06
(3,4)	IP	0	0	0	0	0	0	...	0	0
(3,5)	NI	0	0	0	0	0	0.01	...	0	0.32
(3,5)	IP	0	0	0	0	0	0	...	0	0.02
(4,6)	NI	0	0	0	0	0	0.16	...	0	0.77
(4,6)	IP	0	0	0	0	0	0.02	...	0	0.08
(5,6)	NI	0	0	0	0	0	0	...	0	0.05
(5,6)	IP	0	0	0	0	0	0	...	0	0
(5,7)	NI	0	0	0	0	0	0	...	0	0.29
(5,7)	IP	0	0	0	0	0	0	...	0	0
(6,4)	NI	0	0	0	0	0	0	...	0	0.16
(6,4)	IP	0	0	0	0	0	0	...	0	0.02
(6,8)	NI	0	0	0	0	0	0	...	0	0.68
(6,8)	IP	0	0	0	0	0	0	...	0	0.03
(7,8)	NI	0	0	0	0	0	0	...	0	0.59
(7,8)	IP	0	0	0	0	0	0	...	0	0.03
(7,9)	NI	0	0	0	0	0	0	...	0	0.20
(7,9)	IP	0	0	0	0	0	0	...	0	0
(8,7)	NI	0	0	0	0	0	0	...	0	0.51
(8,7)	IP	0	0	0	0	0	0	...	0	0.03
(8,10)	NI	0	0	0	0	0	0	...	0	0.78
(8,10)	IP	0	0	0	0	0	0	...	0	0.01
(9,T)	NI	0	0	0	0	0	0	...	0	0.2
(10,T)	NI	0	0	0	0	0	0	...	0	0.8

Definition 3.3.1. *(Feasible policy flows.)* A $\left(\sum_{(u,v)\in D} |\Pi_{uv}|\,|Q_{uv}|\right)$-element vector $\mathbf{y}$ of policy flows is *feasible* if $y_q^\pi \geq 0$ for all π and q, and if $\sum_{\pi\in\Pi_{uv}} y_q^\pi = d_{uv}^q$ for all (u,v) and q (that is, $\mathbf{y}$ is nonnegative and satisfies the travel demand.)

These map to state-dependent arc flows x_{ij}^s through application of algorithm LoadPolicy, described in Section 3.2.4. Next, the expected disutility for each policy π connecting user class q_{uv} of OD pair (u,v) can be evaluated as

$$F_q(\pi) \equiv \sum_{t\in T_v} \eta_v^t(\pi) f_q(t) \tag{3.35}$$

The vector of policy disutilities is more compactly written as $\mathbf{F} = [F_q(\pi)]$. With these definitions, the NL-UER condition can be stated as

Definition 3.3.2. *(NL-UER.)* A feasible policy flow vector $\mathbf{y}$ is a nonlinear user equilibrium with recourse (NL-UER) if

$$y_q^\pi > 0 \Rightarrow F_q^\pi = \min_{\pi'\in\Pi_{uv}} \{F_q^{\pi'}\} \tag{3.36}$$

that is, all of the used policies have minimal disutility among the policies available to that user class.

Note that the dependence of $\mathbf{F}$ on $\mathbf{y}$ is suppressed for brevity, but these are related through the mapping of policy flows to arc flows, of arc flows to travel delay, and of travel delay to disutility.

Section 3.3.1 provides an overview of past research in equilibrium with uncertainty, with an emphasis on online equilibrium. Section 3.3.2 briefly considers the case of linear disutility functions, which reduces to Model B of Unnikrishnan (2008). This section also presents an important extension to this work, namely, a more efficient policy loading procedure which also allows consideration of networks with cycles. Treatment of NL-UER begins in earnest with Section 3.3.3, which considers the problem in general and derives conditions for existence and uniqueness of such an equilibrium. Solution methods are presented in Section 3.3.4, with a general method based on variational inequalities. Finally, Section 3.3.5 demonstrates NL-UER on a small network.

3.3.1 Literature Review

The foundations of user equilibrium traffic assignment were laid by Wardrop (1952), in specifying the equilibrium state as one in which no one traveler can reduce his or her travel time by switching routes, and by Beckmann et al. (1956), who show that equilibrium route flows solve a convex optimization program. Fundamental connections also exist between network equilibrium and game theory — for instance, the basic deterministic user equilibrium problem can be viewed as a potential game, a class of economic games with a well-developed theory. This relationship is explored in greater detail in Altman and Wynter (2004).

Incorporating uncertainty in network parameters is somewhat more recent, being performed mainly since the 1980s. Uncertainty in travel demand has been considered by Asakura and Kashiwadani (1991), Bell et al. (1999), Clark and Watling (2005), and Shao et al. (2005), who model day-to-day variations in demand, and by Waller et al. (2001), Ukkusuri and Waller (2004), Duthie (2005), Ukkusuri and Waller (2006), Nagae and Akamatsu (2005), and Lam and Tam (2007) who consider long-term forecasting errors.

In terms of operational uncertainty, a variety of modeling approaches have been proposed. Some researchers have considered network reliability, defined as the probability of two nodes being connected (Iida and Wakabayashi, 1989), or the probability of demand on roadway arcs not exceeding capacity (Chen et al., 2000). Du and Nicholson (1997) and Lo and Tung (2003) model capacity degradations on specific arcs, applying an equilibrium approach to represent the changes in flows and travel times.

Mirchandani and Soroush (1987) developed an equilibrium model accounting both for travel time uncertainty and nonlinear user preferences, a setting which closely resembles that of this dissertation. In a deterministic context, Gabriel and Bernstein (2000) also considered nonlinear user preferences, derived existence and uniqueness conditions, and presented a solution method.

Adaptive routing in equilibrium has not been studied to the same degree as some of the other impacts of uncertainty; much of the work on adaptive routing considers driver response to VMS signs (e.g., Ben-Akiva et al., 1991; Kaysi and Ali, 2000; Tsaavchidis, 2000; Valdez-Diaz et al., 2001), but in a reactive setting where drivers do not anticipate receiving information before choosing their initial routes, and where no equilibration process occurs.

The transit literature has several instances of adaptive routing, to represent phenomena which require users to adopt a contingency route, such as overfull transit vehicles, which require users to adopt a contingency route. Nguyen and Pallottino (1989) and Marcotte and Nguyen (1998) develop equilibrium algorithms based on hyperpaths, which are very similar to the routing policies described in this document. Later, Marcotte et al. (2004) and Hamdouch et al. (2004) show how the hyperpath concept can also be applied in static and dynamic traffic assignment, respectively, to account for drivers who re-route when encountering an overcapacitated arc. Ukkusuri (2005) explores adaptive routing and equilibrium further, showing that information provision may in fact worsen total system conditions, even in equilibrium. Unnikrishnan (2008) made substantial contributions on this topic both theoretically, deriving a convex programming formulation of the problem, and practically, showing how the Frank-Wolfe algorithm (Frank and Wolfe, 1956) can be used to find such an equilibrium on small networks. Gao (2005) also presents a policy-based equilibrium heuristic along with in-depth analysis of the resulting equilibrium state; however, no mathematical properties are derived and the heuristic is not proven to converge.

In this setting, the contribution of this section is the presentation of an equilibrium model accounting simultaneously for operational uncertainty, nonlinear driver behavior (with multiple user classes), and adaptive routing. Previous equilibrium models have only considered a subset of these, so the model presented in this chapter can be seen as a generalization of these. Theoretical properties are presented, along with convergent, exact solution methods.

Furthermore, the equilibrium model most closely resembling this one (Unnikrishnan, 2008) requires an acyclic network, as it is unable to address *contretemps*. From a practical perspective, this assumption is very limiting; in this section, we show how it can be relaxed both in the setting of (Unnikrishnan, 2008), and in the current setting of nonlinear disutility functions.

3.3.2 Extensions for the Linear Case

Before discussing the NL-UER model in its fullness, this section reviews the important special case of linear disutility functions, and shows how to obtain the arc usages $\mathbf{x}$ from policy flows $\mathbf{y}$ in this static setting. As shown in Section 3.2.2, this is equivalent to assuming that all travelers wish to minimize their experienced travel time. That is, we assume that $|Q| = 1$ and $f(t) = t$, and furthermore assume an infinite time horizon $T \to \infty$. In this case, the NL-UER problem reduces to Model B of Unnikrishnan (2008), who proves that the equilibrium policy flows $\mathbf{y}$ solve

$$\min \sum_{(i,j)\in A} \sum_{s\in S_{ij}} \int_0^{x^s_{ij}(\mathbf{y})} c^s_{ij}(x)dx \tag{3.37}$$

where the $x^s_{ij}(\mathbf{y})$ are the state-dependent arc flows generated by the feasible policy flows $\mathbf{y}$ (that is, $\mathbf{y} \geq \mathbf{0}$, and the components of $\mathbf{y}$ associated with an OD pair sum to the corresponding travel demand). Note that the time superscripts are removed, as the disutility functions and policies are stationary in this case. Unnikrishnan (2008) uses an incidence matrix to map each policy to the proportion of its flow which uses arc (i, j) in state s, and then applies the Frank-Wolfe algorithm to solve this program.

One difficulty with this approach is the difficulty of calculating the incidence matrix elements in the presence of possible *contretemps*. Furthermore, as implemented in Unnikrishnan (2008), at each iteration of the Frank-Wolfe algorithm, a temporary matrix of arc states are obtained by loading each OD pair's demand onto the least-cost policy with respect to the current arc costs. In this section, we propose an alternate policy loading mechanism, named LOADPOLICY-STATIC, which can account for *contretemps* and runs more efficiently by loading all policies corresponding to the same destination v simultaneously.

This approach is similar in nature to that of LOADPOLICY, but the static setting eliminates the increasing order-of-time strategy from consideration. The good news is that the state space is much smaller, which more than compensates for the cyclic nature of the graph in typical instances. As before, the algorithm maintains a set of node usages $\boldsymbol{\eta}$ and arc usages $\mathbf{x}$; however, in this case, they are static as well and do not vary with time.

A more critical difference between LOADPOLICY and LOADPOLICY-STATIC is that the node usages are not preserved in the latter; upon termination, $\boldsymbol{\eta} = \mathbf{0}$.

Initially, all of the demand destined for the node v is "loaded" by initializing $\eta_u = d_{uv}$ for all zones u. At each iteration, a node i with positive η_i is then chosen; these vehicles are then divided according to the message θ they receive at i, proportional to the probabilities of each message being received. The vehicles observing each message are then "moved" to the downstream node of the arc (i, j) indicated by the policy, decreasing η_i and increasing both η_j and x^s_{ij} by $\eta_i \rho_\theta$. (This contrasts with LoadPolicy, where the node usages are preserved.) When vehicles arrive at the destination, they are simply removed from the network, so $\eta_v \equiv 0$ at all times.

This method is more efficient because all of the policies corresponding to a single destination are loaded at the same time, eliminating duplication of effort because the optimal policy for one OD pair is also optimal for any other OD pair with the same destination. Typically, policies move vehicles from nodes with higher expected-cost labels $L(\cdot)$ to nodes with lower $L(\cdot)$ (that is, vehicles are usually moved closer to the destination, in an expected sense). Thus, LoadPolicy-Static scans nodes in decreasing order of $L(i)$, and a binary max-heap data structure is used to accomplish this in an efficient manner. Pseudocode for LoadPolicy-Static is shown as Algorithm 3, on page 101.

Substituting this algorithm for the incidence matrix multiplication in Model B of Unnikrishnan (2008) results in an improved algorithm we term UER2.

Contretemps are addressed by specifying a minimum vehicle quota $\eta_{min} \ll ||\mathbf{D}||$ for the number of vehicles η_i at node i. Whenever node i is

Algorithm 3 LOADPOLICY-STATIC(π, v, D)

```
{Arguments: policy $\pi$, destination $v$, OD demand $\mathbf{D}$)}
$\mathbf{x} \leftarrow \mathbf{0}$
$SEL \leftarrow \emptyset$ {Binary max-heap used to identify nodes to scan}
for all $i \in N$ do
  if $i \in Z$ and $D_i > 0$ then
    $\eta_i \leftarrow d_{uv}$
    $SEL \leftarrow SEL \cup i$
  else
    $\eta_i \leftarrow 0$
  end if
end for
while $SEL \neq \emptyset$ do
  Remove a node $i$ from $SEL$ with maximum $L(i)$
  if $\eta_i > \eta_{min}$ then {Number of remaining vehicles is sufficiently large}
    for all $\theta \in \Theta_i$ do
      $(i, j) \leftarrow \pi(i, \theta)$
      for all $s \in S_{ij}$ do
        $x^s_{ij} \leftarrow x^s_{ij} + \rho^\theta_i p^{s\theta}_{ij} \eta_i$
      end for
      if $j \neq v$ then
        $\eta_j \leftarrow \eta_j + \rho^\theta_i \eta_i$
        $SEL \leftarrow SEL \cup j$
      end if
    end for
  else {End *contretemps* by shifting flow to adjacent node with least $L$}
    $j_{min} \leftarrow \arg\min_{j \in \Gamma(i)} \{L(j)\}$
    for all $s \in S_{i,j_{min}}$ do
      $x^s_{(i,j_{min})} \leftarrow x^s_{(i,j_{min})} + p^s_{ij_min}\eta_i$
    end for
    $\eta_{j_{min}} \leftarrow \eta_{j_{min}} + \eta_i$
    $SEL \leftarrow SEL \cup j_{min}$
  end if
  $\eta_i \leftarrow 0$
end while
return $\mathbf{x}$
```

scanned, if $\eta_i < \eta_{min}$, LOADPOLICY-STATIC assumes the presence of a *contretemps*, and interrupts the cycle by moving all η_i vehicles to the adjacent node j with lowest L_j, that is, the adjacent node closest to the destination. Of course, this node may be a part of the cycle defining this *contretemps* as well, in which case the flow may again be shunted onto a neighboring node with lower L, until the cycle is interrupted. This is formalized in the following result, for which two proofs are provided. The first is more instructive, and illustrates how the algorithm handles *contretemps* in practice, while the second is more elegant and, perhaps, more rigorous.

Theorem 3.3.1. *If* π *is an optimal policy, algorithm* FINDPOLICY-STATIC *terminates in finite time.*

Proof. Assume not. Then there is some node i_0 which enters SEL infinitely often. Since $\eta_i = 0$ after i is scanned, there must be some node $i_1 \in \Gamma^{-1}(i_0)$ which is itself scanned infinitely often, and re-enters SEL infinitely often. This logic can be repeated until a cycle $\mathcal{C} = (i_k, i_{k+l} \ldots, i_{k+2}, i_{k+1}, i_k)$ of length l is identified, possibly not involving i_1. Since π is optimal, the probability of any *contretemps* is strictly less than one, and the flow circulating in $\mathcal{C}$ decreases geometrically between successive scans of i_k. Thus, eventually $\eta_{i_k} < \eta_{min}$ and all of η_{i_k} is moved to node $j = \arg\min_{(i_k,j)\in A} L_j$. For the flow to continue to circulate in $\mathcal{C}$, we must have $j = i_{k+l}$ and $L_j < L_{i_k}$, because the positivity of arc travel times implies that each node is incident to another node with strictly lower L. Likewise, for the flow to continue to circulate in $\mathcal{C}$ from i_{k+l}, we require

$L_{i_{k+l-1}} < L_{i_{k+l}}$, and so forth, concluding that $L_{i_k} < L_{i_{k+1}} < \ldots < L_{i_{k+l}} < L_{i_k}$, which is a contradiction. Thus, no cycle is traversed infinitely often, and FINDPOLICY terminates in finitely many iterations. $\square$

Proof. Define the potential function $U = \sum_{i \in N} \eta_i L(i)$. Clearly $U \geq 0$, and $U = 0$ only when no flow remains on the network. We show that whenever a node i is scanned, the change in the potential ΔU is negative and bounded away from zero, which is enough to establish the result.

First, consider the case that $\eta_i > \eta_{min}$. Let (i, j^θ) represent the arc $\pi(i, \theta)$ for all messages θ. In this case,

$$\Delta U = \sum_{\theta \in \Theta_i} \eta_i \rho_i^\theta L(j^\theta) - \eta_i L(i) = \eta_i \left(\sum_{\theta \in \Theta_i} \rho_i^\theta L(j^\theta) - L(i) \right) \tag{3.38}$$

Define $\underline{L} = \min_{i \in N} \left\{ \sum_{\theta \in \Theta_i} \rho_i^\theta L(j^\theta) - L(i) \right\}$. Since arc delays are strictly positive and π is optimal, $\underline{L} < 0$, and thus $\Delta U \leq \eta_i \underline{L} < \eta_{min} \underline{L} < 0$.

When $\eta_i < \eta_{min}$, we also have $\eta_{min} \rho_{min} \leq \eta_i$, where $\rho_{min} = \min\limits_{i \in N, \theta \in \Theta_i} \rho_i^\theta$.[1] Thus $\Delta U \leq \eta_{min} \rho_{min} (L(j_{min}) - L(i))$ which is again bounded away from zero. The theorem immediately follows. $\square$

In practice, η_{min} can be set as small as needed to avoid "false positives" in detecting *contretemps*.

---

[1] This ocurs because, whenever a node i is scanned, at least $\eta_i \rho_{min}$ is added to an adjacent node for each possible message. But if $\eta_i < \eta_{min}$, the entire node usage η_i is added to an adjacent node. Thus, if $\eta_i < \eta_{min} \rho_{min}$, we must have had $\eta_h < \eta_{min} \rho_{min}$ for some $h \in \Gamma^{-1}(i)$ at a previous point in the algorithm. This is seen to be impossible by infinite regress and the algorithm's initialization.

3.3.3 General Properties

In this section, the discussion is broadened to allow nonlinear disutility functions, as well as multiple users classes with differing disutility functions. *Contretemps* are not a significant issue here because the time horizon T automatically forces trips to end in finite time.

Many useful properties of NL-UER can be studied by transforming the problem to an asymmetric static traffic assignment problem, which has been well-studied in the literature. Although this transformation leads to a problem which is far too large to solve directly, it nevertheless is useful for demonstrating equilibrium properties. Construct a graph $G' = (N', A')$ where $N' = \{(u^q, v^q) : (u, v) \in D, q \in Q_{uv}\}$, and where A' contains an arc π^q_{uv} for each OD pair $(u, v) \in D$, user class $q \in Q_{uv}$, and policy $\pi \in \Pi_{uv}$, connecting nodes u^q and v^q. The cost function for each arc π^q_{uv} depends on the flow on each other policy, and denotes the expected disutility for this policy with respect to function f^q. The OD table for G' consists of d^q_{uv} travelers departing u^q for v^q, for all $(u, v) \in D$ and $q \in Q_{uv}$.

Clearly a user equilibrium on G' corresponds to a policy-based equilibrium on G. The following properties of NL-UER thus follow immediately from established results for the asymmetric traffic assignment problem (e.g., Smith, 1979; Dafermos, 1980):

Formulation: Feasible policy flows $\mathbf{y}^*$ are a NL-UER if and only if the variational inequality

$$\mathbf{F}(\mathbf{y}^*) \cdot (\mathbf{y}^* - \mathbf{y}) \leq 0 \tag{3.39}$$

is satisfied for any feasible policy flow vector $\mathbf{y}$.

Existence: If all of the cost functions $t_{ij}^{s}(\cdot)$ and disutility functions $f_{uv}^{q}(\cdot)$ are continuous, a solution to (3.39) exists.

Solution: Any solution algorithm for the variational inequality (3.39) can solve for NL-UER policy flows. If all of the cost functions and disutility functions are monotone as well as continuous, many such algorithms exist.

Uniqueness: If all of the cost functions and disutility functions are continuous and strongly monotone, the solution to (3.39) is unique.

The practical application of this network transformation is limited by its extremely large size: if n, m, Q, S, and T denote the number of nodes, arcs, maximum number of user classes per OD pair, maximum number of states per arc, and latest arrival time in G, G' can include up to Qn^2 nodes and $Qn^2m^{nTS^m}$ arcs.

More tractable solution methods can be drawn from path-based static assignment algorithms, such as those developed by Smith (1983a), Smith (1983b), Lawphongpanich and Hearn (1984), Larsson and Patriksson (1992), and Jayakrishnan et al. (1994). These algorithms do not require enumerating all paths in a network, but only generate paths on an as-needed basis. This results in sets $\hat{\Pi}^q_{uv}$ of "working policies" for OD pair (u, v) and class q, which are much smaller than the sets Π_{uv} of *all* feasible policies, and form the basis of the solution algorithms in the following section.

3.3.4 Solution Algorithms

As is common to many traffic equilibrium algorithms, a three-step iterative procedure can be employed: calculate arc costs from an initial set of flows; determine the shortest paths with respect to these flows; and shift users onto these paths, repeating the procedure with the new set of flows. For NL-UER, these steps can be written as

1. **Policy Evaluation**: Determine the expected disutility for each policy in $\hat{\Pi}^q_{uv}$, $(u, v) \in D$, $q \in Q_{uv}$, using the prevailing policy flows.
2. **Policy Finding**: Given the policy disutility values calculated in the previous step, is there a better policy available for any user class from any OD pair? If so, add it to $\hat{\Pi}^q_{uv}$ for the corresponding user class.
3. **Policy Adjustment**: Given the updated sets $\hat{\Pi}_{uv}$, reassign trips among all of the available working policies.

The first step requires mapping a policy assignment to the expected disutilities of each working policy, which can be accomplished by repeated application of LoadPolicy. This algorithm, EvaluatePolicies, is presented as Algorithm 4; in line 13 of this algorithm, the unit vectors $\mathbf{e_u}$ are given as the OD table argument to LoadPolicy in order to calculate policy disutilities from the node usages.

Algorithm 4 EvaluatePolicies($\mathbf{y}$)

```
 1: {First update arc delays}
 2: $\mathbf{X^t} \leftarrow \mathbf{0}$
 3: for all $(u,v) \in D$, $q \in Q_{uv}$, $\pi \in \hat{\Pi}^q_{uv}$ do
 4:     $(\mathbf{X^t}, \mathbf{H}) \leftarrow (\mathbf{X^t}, \mathbf{H})+$ LoadPolicy($\pi, v, y^q_{uv}\mathbf{D}$)
 5: end for
 6: $\mathbf{X} \leftarrow \sum_t \mathbf{X^t}$
 7: for all $(i,j) \in A$, $s \in S_{ij}$ do
 8:     $t^s_{ij} \leftarrow t^s_{ij}(x^s_{ij})$
 9: end for
10: {Now get policy disutilities}
11: for all $(u,v) \in D$, $q \in Q_{uv}$, $\pi \in \hat{\Pi}^q_{uv}$ do
12:     $\mathbf{H} \leftarrow \mathbf{0}$
13:     $(\mathbf{X}, \mathbf{H}) \leftarrow$LoadPolicy($\pi, v, \mathbf{e_u}$)
14:     $F^q_\pi \leftarrow \sum_{t \in T_v} H(t) f_q(t)$
15: end for
16: return $\mathbf{F}, \mathbf{t}$
```

The second step, policy finding, simply involves executing FindAdaptivePolicy for each OD pair and user class, using the updated travel times found during EvaluatePolicies. Any new policies are added to the working sets $\hat{\Pi}_{uv}$.

The third step requires greater care, as shifting all demand onto the least-disutility policy in the working set is often overcorrection, and may lead to oscillatory behavior in the solution algorithm. It is at this point that techniques from path-based traffic assignment algorithms can be drawn on for assistance, along with the asymmetric transformation in the previous section. For instance, Smith (1983b) shows that the policy shift $\mathbf{y} + \boldsymbol{\Delta}$ moves toward equilibrium, where

$$\boldsymbol{\Delta} = \frac{\displaystyle\sum_{(u,v)\in D}\sum_{q\in Q_{uv}}\sum_{\pi\in\hat{\Pi}_{uv}} [\mathbf{F^q_{uv}}(\mathbf{y})\cdot(\boldsymbol{\Psi_\pi} - \mathbf{y^q_{uv}})]^+(\boldsymbol{\Psi_\pi} - \mathbf{y^q_{uv}})}{\displaystyle\sum_{(u,v)\in D}\sum_{q\in Q_{uv}}\sum_{\pi\in\hat{\Pi}_{uv}} [\mathbf{F^q_{uv}}(\mathbf{y})\cdot(\boldsymbol{\Psi_\pi} - \mathbf{y^q_{uv}})]^+} \tag{3.40}$$

and $\boldsymbol{\Psi_\pi}$ represents an "all-or-nothing" assignment where all d^q_{uv} travelers are assigned to policy π.

An appropriate step size δ is found through calculation of the Smith gap

$$V(\mathbf{F}) = \sum_{(u,v)\in D}\sum_{q\in Q_{uv}}\sum_{\pi\in\hat{\Pi}}([\mathbf{F^q_{uv}}(\mathbf{y})\cdot(\boldsymbol{\Psi_\pi} - \mathbf{y^q_{uv}})]^+)^2 \tag{3.41}$$

and check function

$$W(\mathbf{F}) = \frac{\displaystyle\sum_{(u,v)\in D}\sum_{q\in Q_{uv}}\sum_{\pi\in\hat{\Pi}_{uv}}\sum_{\daleth\in\hat{\Pi}_{uv}} ([\mathbf{F^q_{uv}}(\mathbf{y})\cdot(\boldsymbol{\Psi_\pi} - \mathbf{y^q_{uv}})]^+)^2[\mathbf{F^q_{uv}}(\mathbf{y})\cdot(\boldsymbol{\Psi_\daleth} - \mathbf{y^q_{uv}})]^+}{\displaystyle\sum_{(u,v)\in D}\sum_{q\in Q_{uv}}\sum_{\pi\in\hat{\Pi}_{uv}} [\mathbf{F^q_{uv}}(\mathbf{y})\cdot(\boldsymbol{\Psi_\pi} - \mathbf{y^q_{uv}})]^+} \tag{3.42}$$

Choosing δ so that $V(\mathbf{F}(\mathbf{y} + \delta\boldsymbol{\Delta})) \leq V(\mathbf{F}) - \delta W(\mathbf{F})$ ensures relatively fast convergence.

By contrast, the relative gap

$$\gamma = \frac{\mathbf{F}\cdot\mathbf{y}}{\mathbf{F_{min}} : \mathbf{D}} - 1 \tag{3.43}$$

is more useful for assessing the convergence of the algorithm, irrespective of the magnitude of the disutility functions or the travel demand. (Here $\mathbf{F_{min}}$

is the matrix of minimum-disutility policy costs, and $\mathbf{F_{min}} : \mathbf{D}$ represents the double dot product of this matrix with the OD matrix.) Clearly $\gamma = 0$ if the NL-UER condition is satisfied.

The algorithm then returns to the first step, repeating this process until convergence is obtained.

Putting this all together, algorithm NL-UER (page 110) shows how the steps are combined in order to solve the equilibrium problem. Since this algorithm is essentially solving the asymmetric traffic equilibrium problem on G', convergence to an NL-UER on G is assured.

3.3.5 A Small Demonstration

Demonstrating NL-UER is most clear on a small network; explaining its performance even the simplified Seattle network used in Section 3.2.6 requires a large number of quantities to be simultaneously tracked. Therefore, a smaller example is created, containing four nodes and five arcs (Figure 3.6). All of the arcs have deterministic delay functions except for (2,3), which represents a drawbridge that is either open or closed, with equal probability. When open and traversed by x_{23} vehicles, the delay on the bridge is $4 + 2x_{23}$; when closed, it has infinite cost and is essentially unavailable to travelers.

A total of ten vehicles are traveling from node 1 to node 4, evenly divided into two user classes. The first is risk-neutral, and has the linear disutility function $f^{LINEAR}(t) = t$. The second wishes to maximize the probability of arriving at or before time $t^* = 15$. This would naturally correspond to

Algorithm 5 NL-UER

```
Initialize; get initial flows from free-flow travel times
$x^s_{ij} \leftarrow 0 \qquad \forall (i,j) \in A, s \in S_{ij}$
$t^s_{ij} \leftarrow t^s_{ij}(0) \qquad \forall (i,j) \in A, s \in S_{ij}$
$\hat{\Pi}^q_{uv} \leftarrow$ FindAdaptivePolicy$(\mathbf{t}, f_q, v) \qquad \forall (u,v) \in D, q \in Q_{uv}$
$y^q_\pi \leftarrow d^q_{uv} \qquad \forall (u,v) \in D, q \in Q_{uv}, \pi \in \Pi^q_{uv}$
loop
    {Step 1}
    $(\mathbf{F}, \mathbf{t}) \leftarrow$EvaluatePolicies$(\mathbf{y})$ {Step 2}
    for all $(u,v) \in D, q \in Q_{uv}$ do
        $\pi_{temp} \leftarrow$ FindAdaptivePolicy$(\mathbf{t}, f_q, v)$
        if $\pi_{temp} \notin \hat{\Pi}^q_{uv}$ then
            $\hat{\Pi}^q_{uv} \leftarrow \hat{\Pi}^q_{uv} \cup \pi_{temp}$
            $y_{\pi_{temp}} \leftarrow 0$
        end if
    end for
    if no new policies added then
        return x
    end if
    [Step 3]
    while $\gamma \geq \epsilon$ do
        $\delta \leftarrow 1$
        $\boldsymbol{\Delta} \leftarrow \frac{\sum_{(u,v)\in D} \sum_{q \in Q_{uv}} \sum_{\pi \in \hat{\Pi}_{uv}} [\mathbf{F}(\mathbf{y}) \cdot (\boldsymbol{\Psi_\pi} - \mathbf{y})]^+ (\boldsymbol{\Psi_\pi} - \mathbf{y})}{\sum_{(u,v)\in D} \sum_{q \in Q_{uv}} \sum_{\pi \in \hat{\Pi}_{uv}} [\mathbf{F}(\mathbf{y}) \cdot (\boldsymbol{\Psi_\pi} - \mathbf{y})]^+}$
        while $V(\mathbf{F}(\mathbf{y} + \delta \boldsymbol{\Delta})) > V(\mathbf{F}) - \delta W(\mathbf{F})$ do
            $\delta \leftarrow \delta / 2$
        end while
        $\mathbf{y} \leftarrow \mathbf{y} + \delta \boldsymbol{\Delta}$
    end while
end loop
```

Table 3.6: Policies for NL-UER demonstration

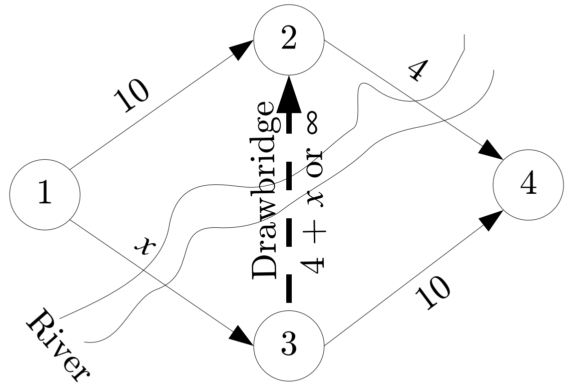
(i, θ)	π_1	π_2	π_3
$(1, \emptyset)$	(1,2)	(1,3)	(1,3)
$(2, \emptyset)$	(2,4)	–	(2,4)
$(3, OPEN)$	–	(3,4)	(3,2)
$(3, CLOSED)$	–	(3,4)	(3,4)

the disutility function $I(t > 15)$; however, this function is discontinuous, and existence of the equilibrium cannot be guaranteed. We modify the function slightly to be continuous, but preserving its general character:

$$f^{AOT}(t) = \begin{cases} 0 & t < 14.5 \\ t - 14.5 & 14.5 \leq t \leq 15.5 \\ 1 & t > 15.5 \end{cases} \tag{3.44}$$

Although eight policies exist in this network, many of them differ only at nodes which are reached with zero probability. Ignoring these, only three distinct policies remain, and are indicated in Table 3.6, where blank entries indicate node states which will never be reached by the policy, and thus the choice of outgoing arc is irrelevant. Descriptively, travelers using policy π_1 will always reach node 4 by traveling on arcs $(1, 2)$ and $(2, 4)$; those using π_2 always use arcs $(1, 3)$ and $(3, 4)$; and those using π_3 will first travel to node 3 and, upon discovering the drawbridge is down, will use the bridge and reach the destination via node 2; if the bridge is up, they will travel directly to the destination.

Table 3.7 shows the progress of the NL-UER algorithm, showing the vector of arc flows $\mathbf{x} = [x_{12}, x_{13}, x_{24}, x_{32}, x_{34}]$, the vectors of policy flows $\mathbf{y^{LIN}} = [y_1^{LIN}, y_2^{LIN}, y_3^{LIN}]$ and $\mathbf{y^{AOT}} = [y_1^{AOT}, y_2^{AOT}, y_3^{AOT}]$ for the two user classes, the



Link delay functions
indicated at left.

Drawbridge is open with
probability 1/2.

Figure 3.6: Demonstration network for NL-UER.

corresponding policy disutilities $\mathbf{F^{LIN}} = [F_1^{LIN}, F_2^{LIN}, F_3^{LIN}]$ and $\mathbf{F^{AOT}} = [F_1^{AOT}, F_2^{AOT}, F_3^{AOT}]$, the Smith gap V, the relative gap γ, and the last step size δ at each iteration. The reader may find it useful to refer to this table in the following discussion.

Initial policies are determined based on free flow travel times. Upon applying FindAdaptivePolicy for both user classes, π_3 is optimal for both, so this policy is added to the working sets Π^{LIN} and Π^{AOT}, and all demand is loaded onto this policy. (The working sets are indicated in the table as well; numerical vector components correspond to working policies, with dashes indicating the remaining policies.) The result of this step is shown as row 0 in Table 3.7.

Continuing, the policies are re-evaluated after the loading, and a new set of optimal policies is determined (row 1-1). Given the prevailing policy usages, π_1 is optimal for both user classes, so this policy is added to the working sets as well.

Proceeding to the equilibration step, the Smith gap and relative gaps are calculated as 2525 and 0.79, respectively. The algorithm will then adjust policy flows until these are in equilibrium before considering the addition of more policies. Improvement directions for both user classes are is given by (3.40) as $\mathbf{\Delta}^{LIN} = [+4.54, -, -4.54]$ and $\mathbf{\Delta}^{AOT} = [+0.45, -, -0.45]$. Using the step size $\delta = 1$ reduces the Smith gap to 12.91, which is sufficient for accepting this move.

Notice that the linear policies are shifted much more than the arriving-on-time policies. This occurs because the difference in the magnitudes of the disutility functions results in the linear policies having disproportional weight in calculating the gap. In practice, the disutility functions should be scaled to be of roughly the same magnitude for common arrival times through suitable affine transformations.

Row 1-2 reflects the new policy flows and disutilities following this adjustment, and new improvement directions $\mathbf{\Delta}^{LIN} = [+0.11, -, -0.11]$ and $\mathbf{\Delta}^{AOT} = [+3.41, -, -3.41]$ are calculated. In this case, the step size $\delta = 1$ is too large (and actually increases the Smith gap). Halving the step size, we see that $V(\mathbf{y} + (1/2)\mathbf{\Delta})$ results in a sufficient gap reduction, and this move is adopted.

The algorithm continues, converging towards the equilibrium policy flows [13/3, –, 2/3] for the linear user class, and [7/3, –, 8/3] for the arriving-on-time user class. Practically, this step is terminated when the gap (relative or Smith) is sufficiently small.

Moving to row 2-1, the algorithm now determines if any new policies need to be added to the working set. For the arriving-on-time class, the two existing working policies suffice. However, for the linear class, policy π_2 has a lower disutility and must be added to the working set. The equilibration process then begins again, now with three policies for the first user class, eventually converging toward the policy flows [1, 2, 2] and [5, –, 0] for the linear and arriving-on-time classes, respectively.

At this point, no new optimal policies can be found, and the algorithm terminates with the equilibrium solution.

3.4 Conclusion

This chapter developed models for individual route choice (NL-OSP) and collective user equilibrium (NL-UER), simultaneously accounting for uncertain travel time, nonlinear disutility functions, and adaptive routing in response to information. These models form the core of the dissertation's contribution: the methods presented in Chapter 2 are primarily aimed at providing the correct parameters for NL-OSP, and NL-UER, and the improvement strategies discussed in Chapter 5 show how these can be applied.

When cycles exist, a traveler's journey may include arbitrarily many cycles, albeit with small probability. For many common disutility functions,

Table 3.7: Demonstration of NL-UER

Step	$\mathbf{x}$	$\mathbf{y}^{\mathbf{LIN}}$	$\mathbf{F}^{\mathbf{LIN}}$	$\mathbf{y}^{\mathbf{AOT}}$	$\mathbf{F}^{\mathbf{AOT}}$	V	γ	δ
0	[0,10.00,5.00,5.00,5.00]	[–,–,5.00]	[–,–,24.00]	[–,–,5.00]	[–,–,1.00]	0	0	–
1-1	[0,10.00,5.00,5.00,5.00]	[0.00,–,5.00]	[14.00,–,24.00]	[0.00,–,5.00]	[0.00,–,1.00]	2525	7.9×10^{-1}	–
1-2	[5.00,5.00,7.50,2.50,2.50]	[4.55,–,0.45]	[14.00,–,16.50]	[0.45,–,4.55]	[0.00,–,0.75]	12.91	6.4×10^{-2}	1
1-3	[6.76,3.24,8.38,1.62,1.62]	[4.60,–,0.40]	[14.00,–,13.86]	[2.16,–,2.84]	[0.00,–,0.00]	0.42	9.4×10^{-3}	$1/2$
1-4	[6.62,3.38,8.31,1.69,1.69]	[4.46,–,0.54]	[14.00,–,14.07]	[2.16,–,2.84]	[0.00,–,0.00]	0.14	5.9×10^{-3}	$1/32$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1-∞	[6.67,3.33,8.33,1.67,1.67]	[4.33,–,0.67]	[14.00,–,14.00]	[2.33,–,2.67]	[0.00,–,0.00]	0	0	0
2-1	[6.67,3.33,8.33,1.67,1.67]	[4.33,0.00,0.67]	[14.00,13.33,14.00]	[2.33,0.00,2.67]	[0.00,0.00,0.00]	26.47	7.2×10^{-2}	–
2-2	[5.92,4.08,7.53,1.61,2.47]	[3.25,0.85,0.90]	[14.00,14.08,14.70]	[2.33,0.00,2.67]	[0.00,0.00,0.41]	2.37	2.3×10^{-2}	$1/4$
2-3	[5.92,4.08,7.30,1.38,2.70]	[3.28,0.90,0.82]	[14.00,14.08,14.47]	[2.64,0.41,1.95]	[0.00,0.00,0.41]	0.80	1.5×10^{-2}	$1/4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2-∞	[6.00,4.00,7.00,1.00,3.00]	[1.00,2.00,1.00]	[14.00,14.00,14.00]	[5.00,0.00,0.00]	[0.00,0.00,0.00]	0	0	0

the probability of this occurrence can be bounded, leading to an upper bound on the error introduced by the time horizon which forces all trips to end.

The NL-UER equilibrium problem was formulated as a variational inequality problem, demonstrating theoretical properties under certain regularity conditions: namely, existence of NL-UER if the delay and disutility functions are continuous, and uniqess if they are monotone. Additionally, an improved policy loading mechanism was found for the (linear) UER problem, allowing all flow destined for the same node to be loaded simultaneously, rather than sequentially. This improved procedure also addresses *contretemps*, and so can be applied in cyclic networks, marking a significant practical enhancement over Unnikrishnan (2008).

Chapter 4

Analysis of Routing and Equilibrium Algorithms

4.1 Introduction

This chapter studies the NL-OSP and NL-UER problems numerically, complementing to the methodological focus of the previous chapter. A suite of test networks was obtained from Bar-Gera (2009), representing standard transportation networks of varying sizes; their properties are shown in Table 4.1. Note that the last two rows of the table gives the average node connectivity and unit of time discretization, respectively.

Section 4.2 focuses on the NL-OSP problem, examining computation time, the impact of different disutility functions, and the impact of varying levels of information provision. Section 4.3 performs a similar analysis for NL-UER. All computation times are obtained from implementing the FIND-ADAPTIVEPOLICY and NLUER algorithms on a 3.4 GHz Pentium 4 machine using Windows XP with 2 GB RAM.

A time discretization of 120 arrival times was used for each network, with the time unit scaled to ensure that both (1) the time horizon was large enough that its impact on the optimal policy is all but nonexistent and (2) the unit was small enough to be commensurable with the lowest arc travel times.

Table 4.1: Characteristics of the test networks

	Sioux Falls	Anaheim	Barcelona	Chicago Regional
n	24	416	1020	12,982
m	76	914	2522	39,018
z	24	38	110	1790
m/n	3.17	2.20	2.47	3.01
Δt	1	1/4	1	1/2

This process resulted in discretizations between 15 and 60 seconds.

Lacking travel data to estimate delay functions for each state, the following structure was used: each arc exists in one of two states, one of which has free flow time equal to that in the deterministic network (occuring with probability 0.9), and the other having free flow time thrice that in the deterministic network (with probability 0.1). Both states share the same capacity.

4.2 Routing Tests

To estimate the average computation time needed by FINDADAPTIVEPOLICY, five OD pairs were randomly generated for each network. These times are reported in Table 4.2. One sees that the run time for each network is very stable across different OD pairs. This should not be surprising, since the decreasing order-of-time label correcting method makes no substantial distinction between different origins and destinations. The variation of run time with network size in practice (as compared to worst-case theoretical bounds) can now be studied.

Since the networks are of comparable connectivity, and have an identical

Table 4.2: Computation time for NL-OSP

Network	OD pair	Time (s)	Average time (s)
Sioux Falls	(1,15)	0.036	0.031
	(3,5)	0.029	
	(2,11)	0.028	
	(7,4)	0.030	
	(12,21)	0.033	
Anaheim	(8,3)	0.44	0.38
	(17,35)	0.38	
	(16,1)	0.36	
	(27,13	0.38	
	(20,4)	0.36	
Barcelona	(104,83)	5.33	5.19
	(22,57)	5.18	
	(5,91)	5.17	
	(50,35)	5.15	
	(5,44)	5.13	
Chicago Regional	(33,1087)	24.22	21.37
	(879,16)	20.75	
	(39,1157)	20.60	
	(1710,1344)	20.65	
	(43,1469)	20.60	

number of states and time steps, the computation time can be related primarily to the number of nodes n. Assuming an exponential relationship

$$T_R = Kn^{\nu} \tag{4.1}$$

between the run time T_R and the network size, with constants K and ν to be determined, a linear regression can be performed between $\log T_R$ and $\log n$ for each of the networks and OD pairs tested above. This results in an estimate of $\nu \approx 1.08$, indicating that the growth in run time is only slightly faster than linear. This is not likely to hold in more networks where the connectivity can grow larger with network size; however, for transportation networks this seems reasonable.

To test the differences in routing policy that occur depending on a user's disutility function, FindAdaptivePolicy was applied to the test networks and OD pairs identified in the previous section. Here, disutility functions requiring a target arrival time (such as deviance, or arriving-on-time) use the expected travel time of the optimal linear disutility solution as the target unless stated otherwise. Boyles (2006) and Boyles and Waller (2007b) studied these differences in a slightly different setting, in random networks and grid networks, respectively, and with a form of limited spatial dependency. This section complements these findings by performing tests in transportation networks based on major metropolitan areas, using the independence assumption which is consistent with the remainder of the algorithms which have been developed. Briefly summarizing, some of the key findings of these past works were

- In random networks, the optimal policies for the linear and deviance disutility functions were identical nearly half the time. When they differed, the minimum-deviance policy involved an increase of 8% in expected travel time and a decrease of 46% in travel time variance, on average.

- The difference between the linear and deviance policies, measured by the mean and variance of expected travel time, was greatest when successive arc costs were positively correlated, and least when negatively correlated. (The differences were intermediate for the case of independent arc costs.)

- In grid networks, monotonic quadratic disutility functions never produced a policy different than what a linear disutility function would have provided, perhaps because the shape of this function is not “twisty” enough in the range of likely arrival times.

- In grid networks, differences between linear disutility functions and the Box-Cox disutility functions typically arose when the shape parameter λ was less than 0.75, or greater than 3. These are slightly outside of the range of λ found by the revealed-preference study by de Lapparent et al. (2002); however, this study focused on mode choice in a deterministic network setting, and it is unclear whether the same λ would apply to route choice in a stochastic environment.

Based on these past findings, three disutility functions are compared: linear, deviance, and arriving-on-time. Tables 4.3, 4.4, and 4.5 show the result

Table 4.3: Results for linear disutility function

Network	OD pair	Linear		
		Mean	Var	Pr(On Time)
Sioux Falls	(1,15)	24.58	9.55	0.82
	(3,5)	7.20	7.20	0.81
	(2,11)	18.78	8.96	0.66
	(7,4)	12.92	9.06	0.66
	(12,21)	11.87	11.25	0.73
Anaheim	(8,3)	22.55	11.43	0.51
	(17,35)	16.27	3.98	0.60
	(16,1)	16.73	4.89	0.59
	(27,13)	14.19	4.00	0.62
	(20,4)	27.09	3.88	0.66
Barcelona	(104,83)	2.98	0.81	0.74
	(22,57)	16.67	3.85	0.57
	(5,44)	10.06	0.36	0.70
	(5,91)	21.48	3.84	0.60
	(50,35)	7.65	0.73	0.61
Chicago Regional	(238,466)	16.34	2.73	0.64
	(510,499)	4.43	0.39	0.84
	(1634,949)	23.2	4.93	0.59
	(33,1087)	51.02	16.51	0.69
	(673,636)	13.08	2.47	0.60

of applying each disutility function to the randomly-chosen OD pairs identified in the previous section, measured by the mean travel time, travel time variance, and the probability of on-time arrival (that is, arriving no later than the mean arrival time).

From these, one can see that the deviance disutility function sharply reduces travel time variance, in accordance with the previous findings in networks of different topology. When resulting in a different policy than a linear

Table 4.4: Results for deviance disutility function

Network	OD pair	Deviance		
		Mean	Var	Pr(On Time)
Sioux Falls	(1,15)	25.88	7.54	0.83
	(3,5)	7.20	7.20	0.81
	(2,11)	18.79	8.94	0.66
	(7,4)	12.92	9.06	0.66
	(12,21)	11.87	11.25	0.73
Anaheim	(8,3)	22.77	10.69	0.52
	(17,35)	17.07	2.00	0.70
	(16,1)	17.48	3.01	0.62
	(27,13)	14.73	3.07	0.65
	(20,4)	27.50	2.99	0.76
Barcelona	(104,83)	3.11	0.79	0.69
	(22,57)	17.42	1.91	0.72
	(5,44)	10.06	0.36	0.70
	(5,91)	22.37	1.82	0.68
	(50,35)	8.05	0.39	0.75
Chicago Regional	(238,466)	16.9	1.37	0.71
	(510,499)	4.43	0.39	0.84
	(1634,949)	24.08	2.57	0.73
	(33,1087)	52.01	12.67	0.76
	(673,636)	13.59	1.42	0.67

Table 4.5: Results for arriving-on-time disutility function

Network	OD pair	Arriving-on-time		
		Mean	Var	Pr(On Time)
Sioux Falls	(1,15)	24.86	12.29	0.82
	(3,5)	7.20	7.20	0.81
	(2,11)	18.78	9.09	0.66
	(7,4)	12.92	9.06	0.66
	(12,21)	11.98	11.77	0.73
Anaheim	(8,3)	22.55	11.43	0.51
	(17,35)	20.96	1.53	0.75
	(16,1)	16.73	4.89	0.59
	(27,13)	14.55	4.10	0.61
	(20,4)	27.50	2.99	0.76
Barcelona	(104,83)	7.73	0.79	0.61
	(22,57)	22.34	0.77	0.82
	(5,44)	16.48	0.29	0.78
	(5,91)	25.08	1.01	0.64
	(50,35)	13.7	0.52	0.60
Chicago Regional	(238,466)	19.14	2.45	0.72
	(510,499)	10.19	0.76	0.67
	(1634,949)	28.95	0.82	0.68
	(33,1087)	51.04	16.28	0.69
	(673,636)	19.84	0.49	0.78

disutility function, a variance reduction of 25–50% is not uncommon, along with a more modest increase in travel time never exceeding 4.5%. This suggests that large gains in reliability are attainable for only a slight increase in expected travel time. Thus, if one believes that travelers are at all sensitive to travel time variability, the risk of misspecifying routing behavior by assuming a linear disutility function is significant.

The arriving-on-time disutility function, on the other hand, results in a policy that very closely resembles that associated with the linear disutility function. Of course, since the “on-time” threshold was specified to be the expected travel time associated with the latter, this connection is not unexpected.

Finally, in practice, information may not be provided at every node. In the tests described in this section, distinct messages can be received at only a limited number of nodes, determined randomly. (Section 5.2 addresses the more interesting problem of identifying the locations where information provision is most beneficial.) Six scenarios were created, with information provided at 0%, 20%, 40%, 60%, 80%, and 100% of the nodes, and a linear disutility function was assumed. In all cases, centroids were never permitted to be information nodes. As can be expected, the expected travel time decreases with more information provided, as shown in Figure 4.1 for the Sioux Falls network. Similar trends are observed for the other networks as well.

A more interesting relation is the dependence of the run time on the amount of information provided. Figures 4.2-4.2 plot this relationship, and

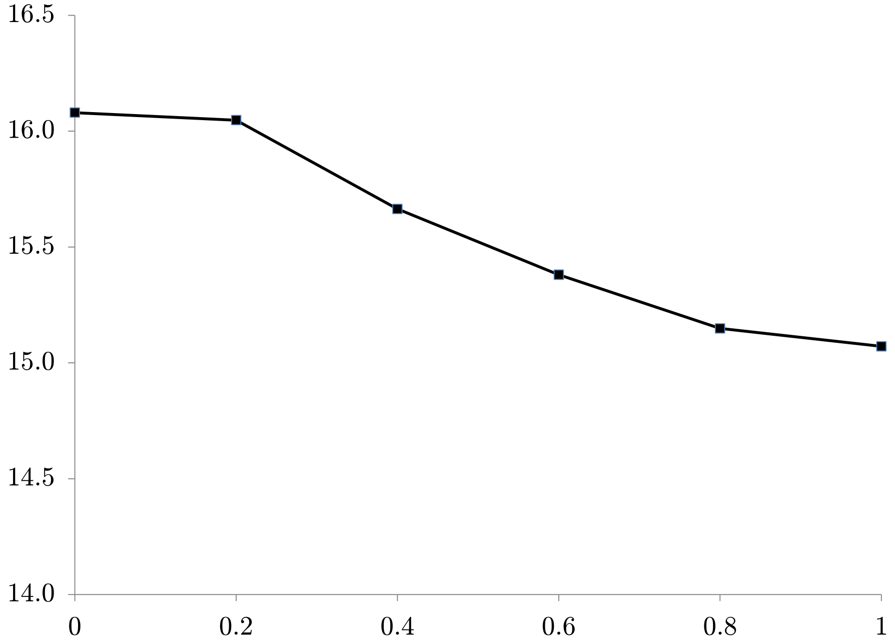


Figure 4.1: Expected policy cost with varying information levels, Sioux Falls.

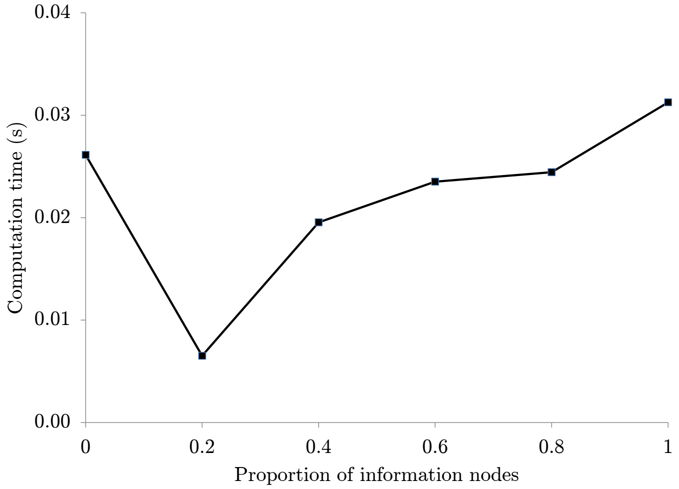


Figure 4.2: Computation time with information level, Sioux Falls.

provide several insights. Generally speaking, the larger the network, the greater the decrease in run time as the number of information nodes shrinks. For instance, in Sioux Falls, the variation in run times is nonuniform and an increase is barely visible, while a steady, near-linear relation is clear for Chicago. This is sensical: on smaller networks, the proportion of total run time consumed by initialization and other "overhead" is much greater, so the savings from easier computation of node labels is reduced. From the standpoint of practice, this trait is desirable, since the time savings obtainable by restricting the information locations (e.g., to freeways and major arterial intersections) actually increases for larger networks.

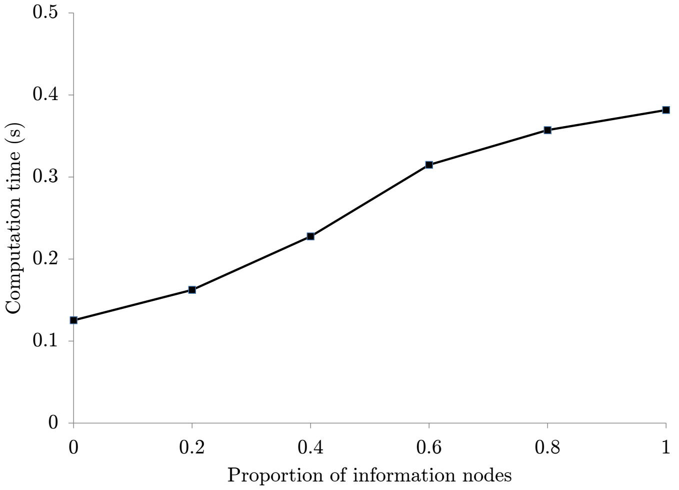


Figure 4.3: Computation time with information level, Anaheim.

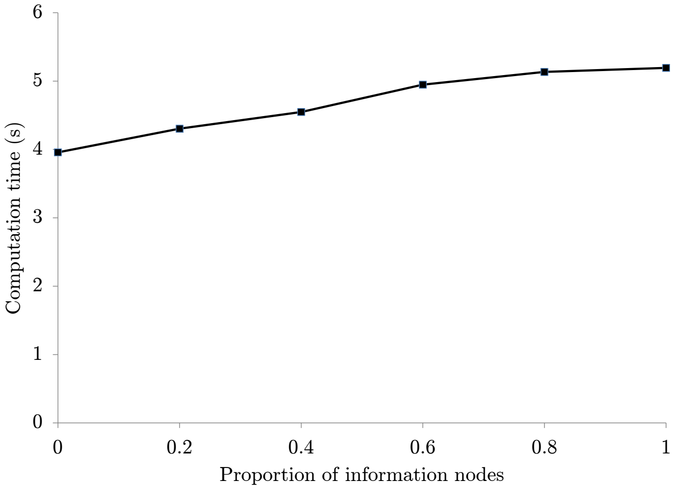


Figure 4.4: Computation time with information level, Barcelona.

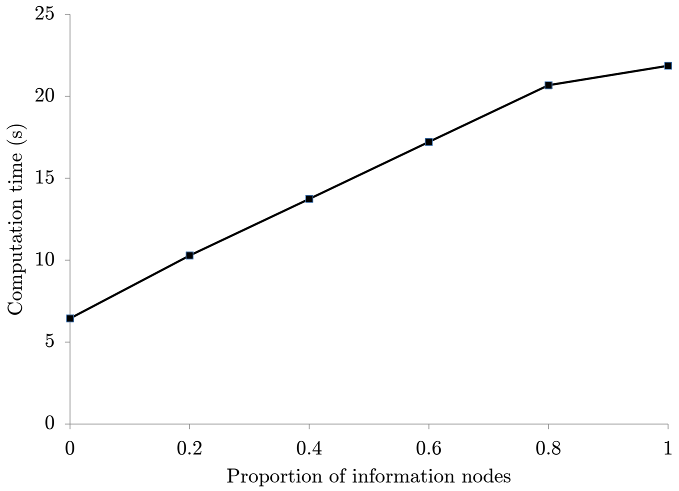


Figure 4.5: Computation time with information level, Chicago Regional.

4.3 Equilibrium Tests

The NLUER algorithm was applied to the Sioux Falls network in order to assess its performance, to compare the impact of different disutility functions, and to examine the benefits of information. None of the larger networks were able to be tested due to memory reasons. However, note that Sioux Falls was also the largest network tested for the (implicitly risk-neutral) UER algorithm (Unnikrishnan, 2008). As the time discretization increases the state space by a factor of 120, the NL-UER algorithm still represents a significant advance in efficiency.

The rate of convergence was measured by plotting the relative gap γ measured over time (in seconds), as shown in Figures 4.6 to 4.8 for three dif-

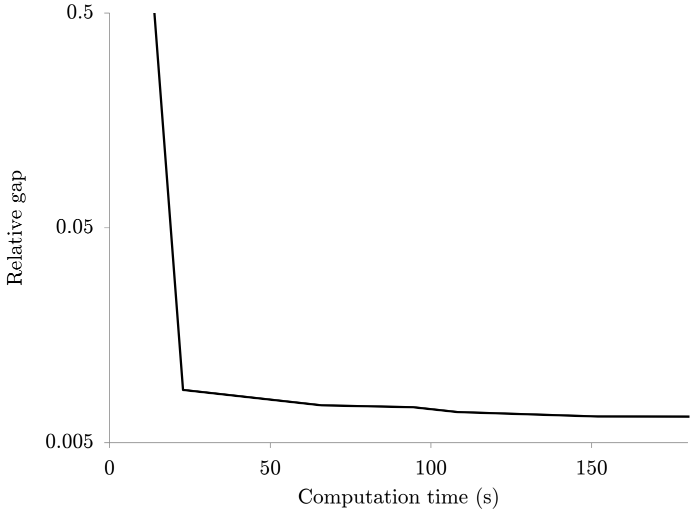


Figure 4.6: Convergence of NL-UER for linear-disutility travelers.

ferent scenarios representing all travelers having a linear disutility function, all travelers having the quadratic disutility function $f(t) = t^2$, and a mixture where half of travelers' disutility is linear, and half is quadratic. As is common to all traffic equilibrium algorithms, initial reduction in the gap is rapid, with final convergence to equilibrium much slower. In particular, the scenario with two user classes is slower to converge because of the additional work in determining optimal policies for the second user class, as well as loading and equilibrating among these policies. As the algorithm design suggests, and as Figures 4.6–4.8 confirm, this additional time is roughly linear in the number of user classes.

Figure 4.9 compares the total system travel time (TSTT) of the best-

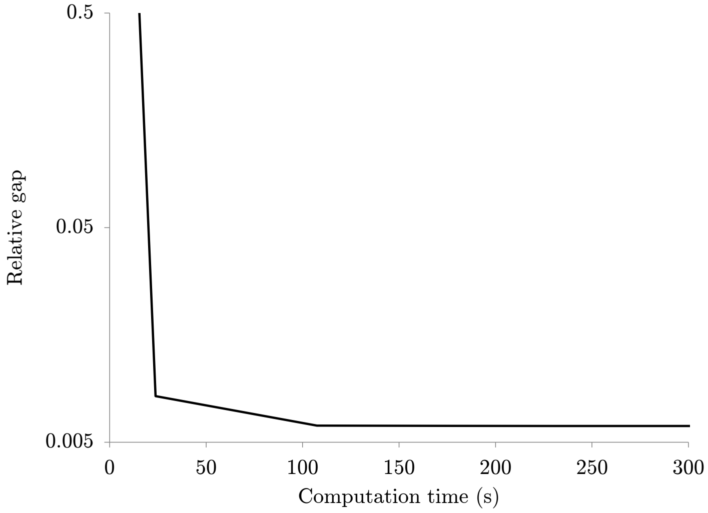


Figure 4.7: Convergence of NL-UER for quadratic-disutility travelers.

converged solutions for each user class scenario, with the total system travel time defined as

$$TSTT = \sum_{(i,j)\in A} \sum_{s\in S_{ij}} x_{ij}^s t_{ij}^s(x_{ij}^s) \tag{4.2}$$

Again, note that this value is *deterministic*. Although individuals perceive travel times stochastically, the assumptions of independently-observed states and a continuum of infinitesimal users result in a deterministic aggregate system state. Further, the travel time is measured, rather than the disutility, in order to reflect an observable quantity and to facilitate comparison among the different scenarios.

As may be expected, TSTT is lower for the linear disutility functions

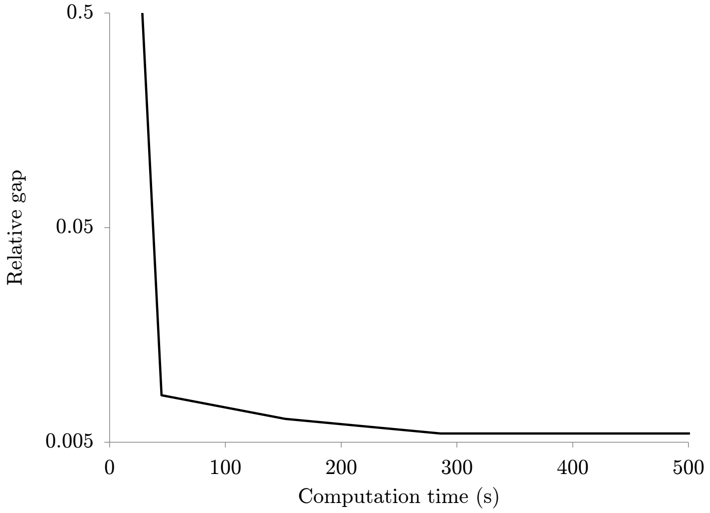


Figure 4.8: Convergence of NL-UER for mixed travelers.

than for the quadratic disutility functions, which reflect risk aversion. Notice that this risk aversion occurs in spite of the deterministic nature of the aggregate flows x_{ij}^s, because each individual's decision reflects uncertainty in arc states. More surprisingly, when the demand is split among two user classes, the TSTT is lower than that observed for the two classes individually. The distinction between user equilibrium and system optimum flows can help interpret this result. In traffic equilibrium problems, aggregate system conditions can be improved if selected individuals move from faster routes with a high marginal travel time to slower routes with a low marginal travel time. This is exactly the behavior that occurs if part of the population is risk averse (for instance, moving from a congestible freeway to a longer arterial). Of course,

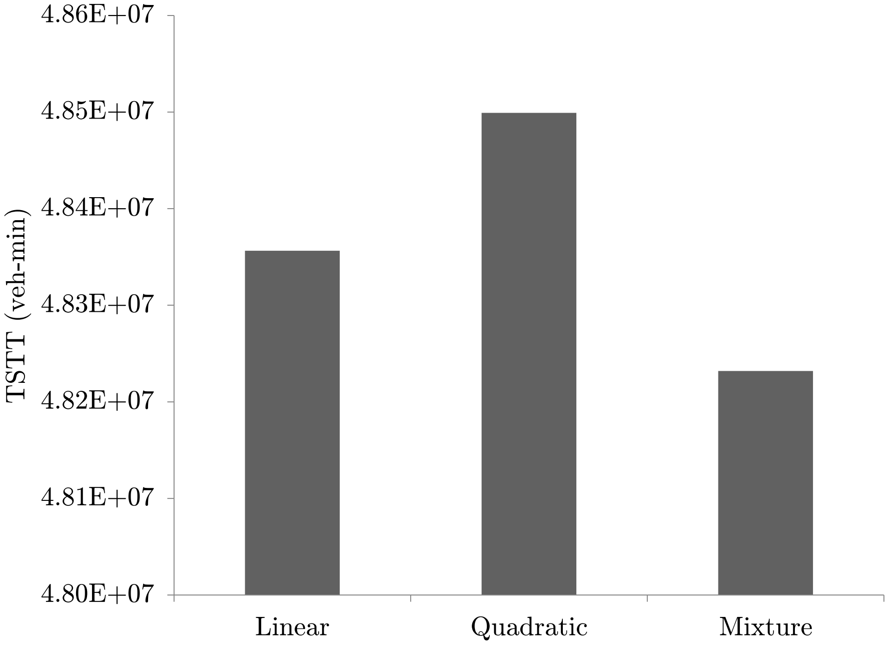


Figure 4.9: Comparing user class scenarios for NL-UER

if too many travelers behave in this way, system conditions can worsen, as seen by the higher TSTT for the scenario where all travelers exhibit quadratic disutility functions.

Finally, the impact of information provision is considered in Figure 4.10; here, all travelers are assumed to have linear disutility. As with NL-OSP, six scenarios are considered, with 0%, 20%, 40%, 60%, 80%, and 100% information provision, with the information nodes selected randomly among non-centroids. In this network, when more information is provided, TSTT is lower. This property is intuitive, although not guaranteed; as shown in Ukkusuri (2005), providing more information to self-optimizing travelers need not improve aggregate system conditions. However, this effect, reminiscent of Braess' paradox (Braess, 1969), is certainly not universal and is not observed here.

Interestingly, most of the benefits of added information seem to accrue when the number of information nodes is already large (80% – 100%). This appears to be an artifact of the random distribution of assignment nodes; as seen in Section 5.2, when information nodes are chosen *optimally*, rather than randomly, a large portion of the benefits of information can be obtained even when provided only at a few nodes.

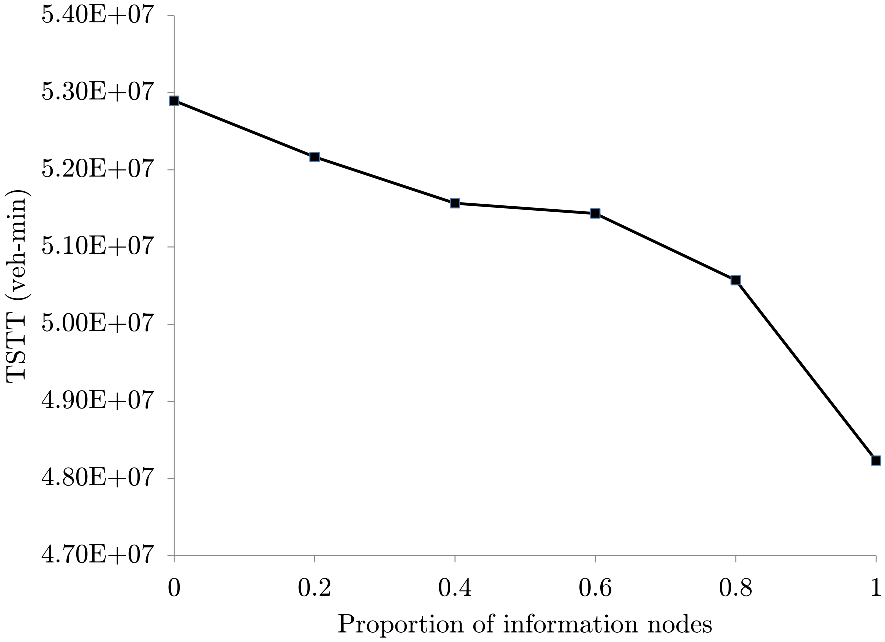


Figure 4.10: The impact of information provision on NL-UER

Chapter 5

Improvement Strategies

5.1 Introduction

Describing user behavior through routing and equilibrium is generally not useful in and of itself. Rather, the purpose of these models is to inform decisions for improving transportation networks, by quantifying the problems currently exist, and by predicting the impact that any improvement strategy might have.

The spectrum of improvement strategies is vast, but three potential strategies are selected for discussion in this chapter: information provision, congestion pricing, and network design. Information provision concerns the location of devices such as VMSs or highway advisory radio (and, interestingly enough, can also be applied to produce adaptive driving directions to individuals before departing), and is impossible to study without some form of adaptive routing and equilibrium model. Congestion pricing is a more familiar strategy, but the possibility of dynamic pricing is newer, and again demands an adaptive behavior model to fully represent driver actions. Finally, the transportation network design problem is concerned with identifying the locations where capacity improvements provide the most improvement in system conditions.

5.2 Information Provision

One common strategy for mitigating uncertainty is information provision through advanced traveler information systems (ATIS), such as VMS signs, highway advisory radio (HAR), or many other technologies. These devices often provide information to drivers *en route*, so while drivers may anticipate receiving information at certain locations, they cannot anticipate the specific message they will receive. Thus, adaptive routing algorithms are needed to describe how drivers respond to this type of information.

Within this context, public agencies must make decisions about where to locate devices such as VMSs or HARs. Installing these devices is costly, and a limited budget is available — for instance, an agency may only have sufficient funds for placing three VMS signs in a certain city, and must decide how to locate them to maximize the benefit to drivers.

Alternately, the information location problems can also be used to provide adaptive driving directions for individuals. Many services are available which provide a route connecting a given origin and a given destination; however, in congested regions, the expected travel time can be reduced by providing several alternatives which can be used depending on observed traffic conditions. Online shortest path algorithms can provide some insight on this problem, but their practical application is limited to real-time devices (such as in-vehicle navigation systems) because these typically assume a re-routing decision can be made at *every* node, and there is no easy way to convey this to drivers through printable directions or other format given *a priori.* On the other hand, by restricting re-routing decisions to a small number of nodes, one can simply report several complete paths to drivers, which is far more easily understood — the problem becomes one of deciding where to allow this re-routing, which is identical to the VMS location problem faced by a public agency. In this case, it may not be necessary to assume an external information provision device, but base online decisions on qualitative observations made by the driver: “If the freeway is congested, exit onto this arterial.”

Several researchers have conducted studies regarding optimal locations for providing information. Abbas and McCoy (1999) applied a genetic algorithm to place VMSs at locations that maximize the number of vehicles which observe these signs, but did not consider adaptive behavior in response to this information. Chiu et al. (2001) and Chiu and Huynh (2007) combine a mesoscopic dynamic traffic assignment simulation with a tabu search heuristic to optimally locate VMSs. Incidents were randomly generated using a Monte Carlo scheme, and some drivers would switch routes if their path encounters an incident and a VMS sign; based on the resulting flow patterns, a set of VMS locations was determined to optimize some measure of effectiveness. Huynh et al. (2003) uses a similar analysis framework to find the optimal locations of portable VMSs in a real-time framework, using the G-D heuristic. Although the simulation approach allows a rich set of traffic and behavioral impacts to be modeled, the computational burden associated with many simulation runs on a large network can be troublesome.

This limitation was realized by Henderson (2004), who adopted a static equilibrium framework for VMS location, together with a discrete choice model to determine the proportion of drivers who switch routes in response to learning of an incident. Several heuristic techniques are developed and compared, including a genetic algorithm and a greedy approach based on sequential location. While compuationally faster, this approach implicitly assumes that drivers do not anticipate receiving information; that is, their initial route choice is not affected by the VMS locations, so arcs with a VMS do not "attract" drivers who anticipate benefitting from that information, for instance. Although this distinction may seem subtle, this anticipation effect can lead to radically different route choices for rational drivers, even from the origin (Boyles, 2006).

The research presented here complements these works by providing analytical network algorithms for locating information, where users both anticipate receiving information and adjust their routes adaptively. The remainder of this section is organized as follows. Section 5.2.1 describes the problem context formally, along with rigorous definitions of three information location problems addressed here. Section 5.2.2 describes a network contraction procedure which allows candidate solutions to be evaluated extremely rapidly. Section 5.2.3 describes exact algorithms and heuristics for solving these three problems, which are then demonstrated in Section 5.2.4.

5.2.1 Problem Definitions

Recall that drivers receive travel information at a set of *information nodes* $R \subset N$. For simplicity, assume that all travelers have the linear disutility function $f(t) = t$, and that only two types of message structures are allowed: full information, where $\theta_i \in \mathcal{S}_i$, and no information, where $\theta_i = \mathcal{S}_i$. It is sometimes more convenient, and perhaps more telling, to express the "no information" message as $\emptyset$.[1] Note that the set of information nodes is a decision variable in information location problems, whereas previously they were exogenous.

For instance, consider the network shown in Figure 5.1, where the arc labels represent delays. Arcs (A, B), (C, E), and (D, E) have deterministic travel time, while the delays on (A, E), (B, C) and (B, D) take on one of two values with equal probability. Assume that there is one traveler departing node A and destined for node E. If node A is an information node, the traveler learns the travel time on (A, E) and (A, B), so the potential messages are $\{(7, 2)\}$ and $\{(8, 2)\}$ and $\Theta_A = \{\{(7, 2)\}, \{(8, 2)\}\}$. On the other hand, if A was not an information node, $\Theta_A = \{\{(7, 2), (8, 2)\}\}$, and the driver must choose a path without knowing the exact delays on these arcs because the only message $\{(7, 2), (8, 2)\}$ (read: "Arc (A, E) either has travel time 7 or 8, arc (A, B) has travel time 2") tells nothing. Likewise, if node B were an information node, the messages would indicate the delays on (B, C) and

---

[1] The curious duality of representing "no information" by both $\emptyset$ and the full joint state $\mathcal{S}_i$ reflects that, by saying all joint states are possible, the message essentially conveys nothing.

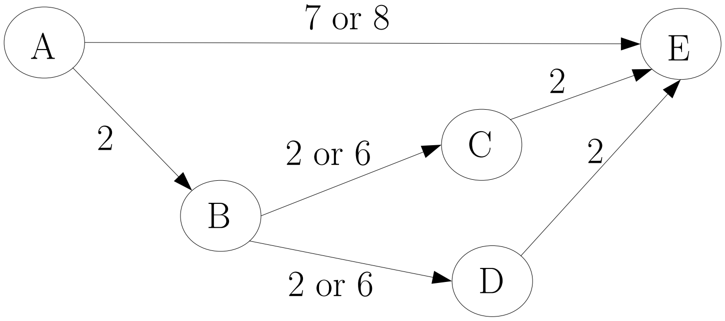


Figure 5.1: Example network to demonstrate notation and concepts.

(B, D), with $\Theta_B = \{(2,2),(2,6),(6,2),(6,6)\}$.

Continuing, consider the case where $R = \{B\}$, that is, B is the only information node. The set of node-states is shown in Table 5.1, along with the least-expected time routing policy. (To simplify matters, in this section we assume that travelers have linear disutility functions. Allowing nonlinear disutility functions is not difficult mathematically, but greatly complicates the notation, primarily by removing the time dependence from problem.)

Since node A is not an information node, the driver will always choose to travel to node B, at which point the delays on (B, C) and (B, D) will be revealed. If either of these arcs is in the "low" state (travel time 2), it will be chosen by the driver, who will then continue on to node E and experience a total travel time of 6 units. The only way an arc will be traversed in the "high"state (travel time 6) is if both (B, C) and (B, D) have high travel time, which occurs with probability $1/4$ and results in a total travel time of 10 units. Thus, the expected disutility of this policy is $6 \times 3/4 + 10 \times 1/4 = 7$ units.

Table 5.1: Node states and optimal policy for the example network with $R = \{B\}$

Node state	Chosen arc
$(A, \emptyset)$	(A, B)
$(B, (2, 2))$	(B, C)
$(B, (2, 6))$	(B, C)
$(B, (6, 2))$	(B, D)
$(B, (6, 6))$	(B, C)
$(C, \emptyset)$	(C, E)
$(D, \emptyset)$	(D, E)

Note that the driver exhibits anticipatory behavior: the only reason for traveling to node B is because information will be revealed at that point. Without information and adaptive routing, the least expected-time path is simply to follow arc (A, E) directly to the destination, with expected travel time 7.5; this demonstrates that the driver's route choice at the origin can be affected by information provided at a later time.

Table 5.2 shows the set of node states and optimal policy if A was the only information node, rather than B. In this case, the optimal strategy is to always choose arc (A, E), with an expected travel time of 7.5. Therefore, in this example, it is better to provide information at node B rather than node A, because the resulting optimal policy has lower expected travel time.

Let the cost of providing information at node i be given by C_i, and assume that a given budget B is available for this purpose. These costs can either be monetary (as with a public agency seeking to install VMS signs) or abstract (as with driving directions, where one can use unit cost for C_i and set

Table 5.2: Node states and optimal policy for the example network with $R = \{A\}$

Node state	Chosen arc
$(A, (7, 2))$	(A, E)
$(A, (8, 2))$	(A, E)
$(B, \emptyset)$	(B, C)
$(C, \emptyset)$	(C, E)
$(D, \emptyset)$	(D, E)

B to the maximum number of information nodes). Within these assumptions, we consider three different information location problems. In each case, the goal is to find a set of information nodes $R^* \in \overline{R}$ optimizing a particular objective, where $\overline{R}$ represents the set of feasible information node sets (that is, the information node sets whose cost does not exceed the available budget).

Individual Information Provision (IIP) In this problem, we are only optimizing a single traveler's expected travel time, so only one element of $\mathbf{D}$ is nonzero, and the delay functions t^s_{ij} are constant, because an atomic individual's travel decision will not affect the travel times they experience. This problem is appropriate for providing adaptive driving directions for an individual with a private service.

Uncongested Information Provision (UIP) In this problem, we are concerned with minimizing the total system travel time of a large number of travelers, where congestion effects are ignored:

$$TSTT = \sum_{(i,j) \in A} \sum_{s \in S_{ij}} x^s_{ij} t^s_{ij} \tag{5.1}$$

Table 5.3: Overview of problems IIP, UIP, and CIP

	IIP	UIP	CIP
OD pairs	One	Many	Many
Arc delays	Constant	Constant	Flow-dependent
Objective function	$F(\pi^*)$	$TSTT$	$TSTT$
Key algorithm (linear)	TD-OSP	TD-OSP	UER2
Key algorithm (nonlinear)	NL-OSP	NL-OSP	NL-UER

That is, **d** may take on general values, but the delay functions t^s_{ij} are still constant. This problem is appropriate for representing information provision on large networks with minimal congestion, such as freight routes in rural areas where weather closures may require re-routing.

Congested Information Provision (CIP) In this case, we are again concerned with minimizing the total system travel time, but here congestion effects must be considered, so the delays t^s_{ij} will depend on the flows x^s_{ij}. This is appropriate for representing urban areas where incidents cause significant reliability issues.

Clearly, IIP is a special case of UIP, and both of these are special cases of CIP. Table 5.3 briefly summarizes the differences between these problems, where the first algorithm listed is used for nonlinear disutility functions, and the second for linear ones. TD-OSP is described in Waller and Ziliaskopoulos (2002), and the reader is referred there for further details.

Finally, as a practical note, it is well-known that not all drivers will switch routes in response to information received *en route*. For the purposes of information location, such users can be ignored as long as the number of such drivers is known, by incorporating their presence into the delay functions as "background" traffic. Behavioral models where switching occurs only under certain circumstances (trip purpose, degree of time savings, freeway vs. arterial) are not considered in the present work.

5.2.2 Network Contraction

It is not trivial to evaluate a given set of information nodes R. The most straightforward approach is to apply an online routing or equilibrium algorithm to the network with information nodes R. Assuming that disutility functions are linear, for IIP this consists of a single application of TD-OSP to determine the expected travel disutility from the origin to the destination with R the information nodes. For UIP, because TD-OSP calculates an "all-to-one" optimal policy tree, one can calculate the total system travel time by applying TD-OSP n times, once for each possible destination, multiplying the expected travel disutility from each origin by the travel demand, and summing over all origins and destinations. For CIP, the UER algorithm must be run to convergence.

This direct approach is undesirable for two reasons. First, applying these algorithms requires some computation time, and any conceivable solution algorithm requires evaluation of a large number of potential information node sets. Second, the computation time required for each of these algorithms grows with network size: TD-OSP requires $O(n^2 mS \log(nS))$ time, where $S = \max_{ij} |S_{ij}|$, and UER2, which involves repeated solution of TD-OSP, exhibits comparable growth in run time.

The good news is that a faster approach for evaluating information nodes is available for IIP and UIP, allowing TD-OSP to be applied to a much smaller network. For simplicity, we first describe this procedure for IIP, then show how it is adapted for UIP.

Because drivers can only make a recourse decision at an information node, their routes they travel are deterministic except at such nodes, simply because they do not receive any information which would cause them to switch paths. Furthermore, at information nodes, drivers only learn information about adjacent arcs. Upon arriving at the downstream end of these arcs, they will continue to follow a deterministic path until encountering another information node or the destination.

This can be represented by constructing a contracted network $G^C(R) = (N^C(R), A^C(R))$, where the contracted node set $N^C(R)$ consists of the origin, the destination, the information nodes R, and the nodes adjacent to information nodes, and where the contracted arc set $A^C(R)$ connects the origin to each information node and the destination, each information node to its adjacent nodes, and every adjacent node to each information node and the destination. Figure 5.2 shows a sample contracted network for two information nodes (marked in grey). In this figure, solid lines represent arcs which also exist in the original network G ("direct arcs"), while dashed lines represent a deterministic path connecting its tail and head nodes in G ("path arcs"). The only direct arcs are those connecting recourse nodes to their

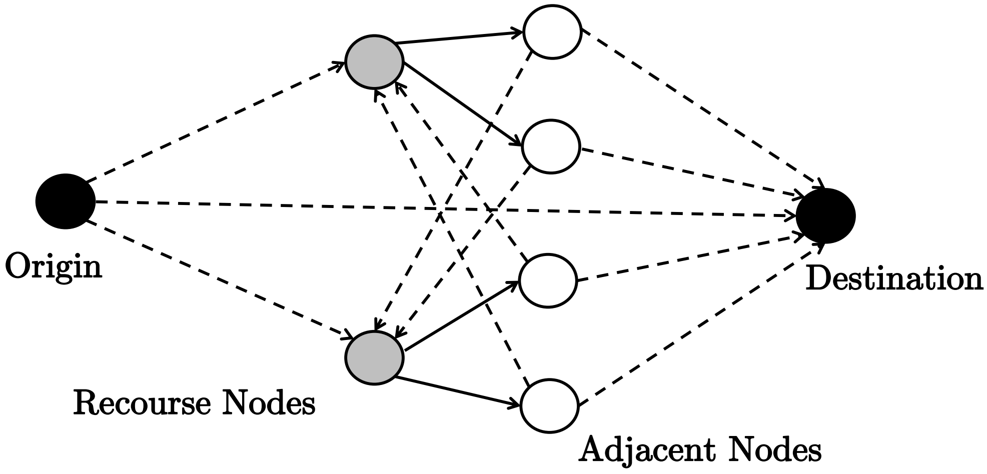


Figure 5.2: Example contracted network with $k = 2$.

adjacent nodes; all of the other contracted arcs represent paths in G. We denote the set of direct and path arcs as $A^C_D(R)$ and $A^C_P(R)$, respectively. Note that $N^C(R)$ contains $2 + |R| + \sum_{i \in R} |\Gamma(i)|$ nodes and $A^C(R)$ contains $(1 + \sum_{i \in R} |\Gamma(i)|)(|R| + 1) + \sum_{i \in R} |\Gamma(i)|$ arcs.

To demonstrate this concept using a larger network, Figure 5.3 shows how a contracted network is created on the Sioux Falls network. The black nodes denote the origin and the destination, while the grey nodes indicate the information nodes. Of the two travel times shown in the original network, the lower one occurs with probability 0.9, while the higher one occurs with probability 0.1.

Note that the only arcs in the contracted graph with uncertain travel times are those adjacent to information nodes, since these are the only locations

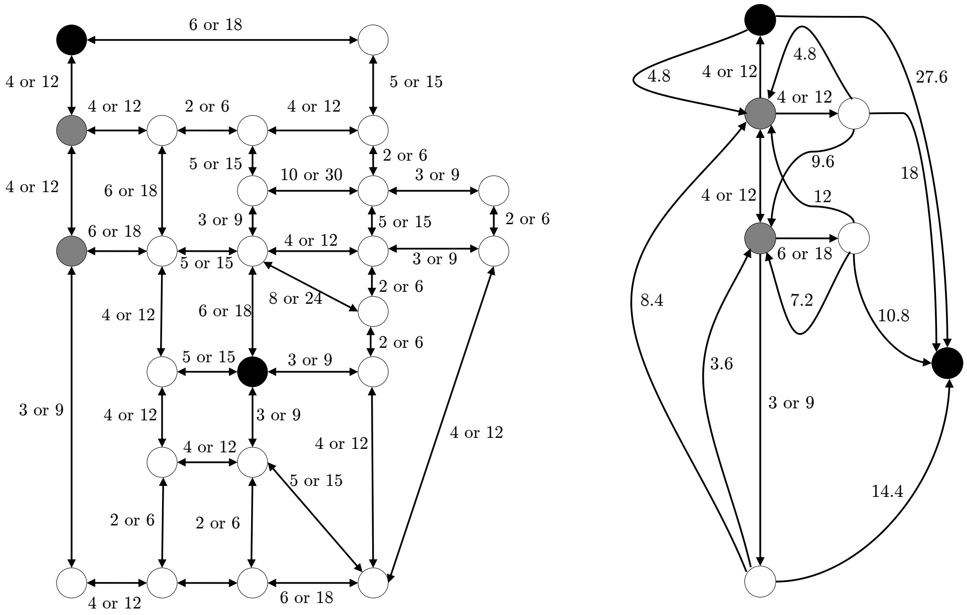


Figure 5.3: Sioux Falls network and a contracted graph for two information nodes.

where an adaptive decision can be made. The remaining nodes are connected by arcs with deterministic travel time, representing the least expected-time path between these. (The justification for choosing these delays is given in Theorem 5.2.1.) Although this network is only slightly smaller than the original network, the contracted network would be nearly the same size regardless of the number of nodes and arcs in the original graph, assuming the node connectivity is comparable.

In particular, by choosing the path arcs to represent least expected-time paths between their tail and head nodes, and by setting the arc's delay to the expected travel time of this path, the optimal policy π^C on the contracted graph has the same disutility as the optimal policy π^* on the original graph, as shown below.

Theorem 5.2.1. $F(\pi^C) = F(\pi^*)$

Proof. We first show that $F(\pi^C) \leq F(\pi^*)$. Consider the following procedure CONTRACT, applied to a node i where $\Theta_i = \{\emptyset\}$: eliminate i from the graph, along with all arcs adjacent to i. For each arc incident to i, replace that arc's head node with $\pi^*(i, \emptyset)$, and add the expected travel time of arc $(i, \pi^*(i, \emptyset))$ to the delay of each of its states. Note that the disutility of π^* is unaffected by this procedure (in fact, the policy itself is essentially unaffected, aside from the trivial removal of node-state $(i, \emptyset)$). Returning to graph G, iteratively apply CONTRACT, each time choosing a node i which is neither an information node, nor immediately adjacdent to an information node. Each step does not affect the disutility of the optimal policy, and the resulting graph is a subgraph of G^C (since clearly the deterministic components of π^* must represent least expected-time paths), implying $F(\pi^C) \leq F(\pi^*)$.

Similarly, we can show that $F(\pi^*) \leq F(\pi^C)$, which is enough to prove the result. Since the arcs A_P^C represent least expected-time paths in G, a policy in G with equal expected disutility can be trivially constructed by expanding the policy π^C using these paths, unless there exists a node $j \in N$ which is part of two such shortest paths to different nodes k and l (see Figure 5.4(a)). Thus, assume that such a node exists.[2] Let L_k and L_l be labels representing

---

[2] Essentially, at non-information nodes, a traveler following a policy in G must make the

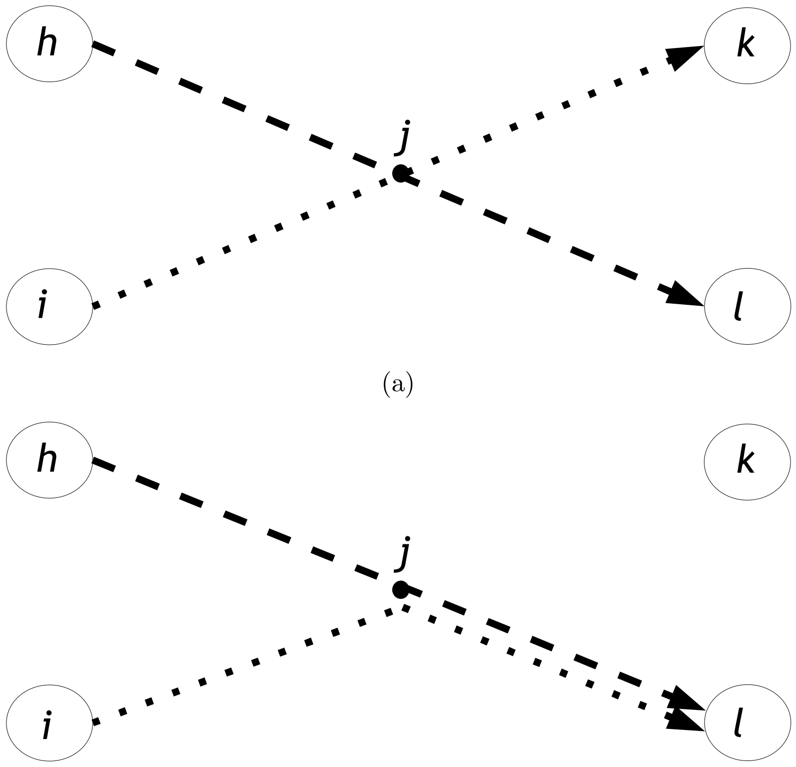
the expected travel disutility from k to the desination v; since arc states are independent, these labels do not depend on the path taken to reach these nodes. Since π^C is optimal, $L_k \leq L_l$, because otherwise the path segment $j-k$ could be replaced by $j-l$. By the same argument, $L_l \leq L_k$ and thus $L_l = L_k$. Thus, when constructing a policy in G from π^C, altering one of the expanded paths from a path arc to be consistent with the expanded path from another (Figure 5.4(b)) does not change the expected disutility of the policy. □

The contracted graph is extremely useful for solving IIP and UIP because it allows the value of a set of informationnodes to be evaluated by applying TD-OSP to a much smaller graph. In particular, note that the size of the contracted graph does not depend on the size of the original graph. Since $|R| \ll n$ in most cases, this leads to an enormous reduction in the time needed for evaluation. (Of course, the number of feasible sets of information nodes still grows with network size.)

To evaluate a set of information nodes for UIP, one might imagine that a contracted graph should be constructed for each OD pair in Z^2. However, a more efficient approach is possible. Because TD-OSP is an "all-to-one" label correcting algorithm, it suffices to construct a single contracted graph for each destination, provided that every origin node is included as well; that is, taking

---

same decision regardless of their past travel history, while a traveler following a path arc has an additional piece of information — the tail and head nodes of that path. We must show that this additional information cannot improve the expected disutility of the optimal policy.



(a)

(b)

Figure 5.4: Potential conflict with expanding policies on the contracted graph, and resolution procedure.

the union of all of the contracted graphs corresponding to a single destination. The contracted graphs formed in this manner will contain $1 + |Z| + |R| + \sum_{i \in R} |\Gamma(i)|$ nodes and $(|Z| + \sum_{i \in R} |\Gamma(i)|)(|R| + 1) + \sum_{i \in R} |\Gamma(i)|$ arcs.

One might object that performing TD-OSP $|Z|$ times to the slightly larger destination-based networks is worse than $|Z|^2$ applications on the smaller single origin-destination networks, because TD-OSP grows faster than linearly in network size. However, since the origin nodes have no reverse star, their addition involves very little increase in the run time, certainly much less than the worst-case bound. A better comparison is the number of node labels which must be calculated; with the given graph sizes, using the destination-based networks requires the calculation of $(|Z|^2 - |Z|)(1 + \sum_{i \in R} |\Gamma(i)|)$ fewer labels than the use of the single origin-destination networks, a savings which is substantial in large networks where many network contractions need to be performed.

Unfortunately, this contraction procedure is not useful for CIP, because the arc delays are flow-dependent, implying that multiple paths will be used by each OD pair in general, and thus generating the appropriate delay function for the path arcs is difficult.

5.2.3 Solution Methods

All three of the information location problems described above are difficult to solve exactly, as IIP, UIP, and CIP are essentially facility location or network design problems, where the solution cost is determined by the travel time experienced by the driver(s). Such problems are notoriously difficult to solve due to their nonlinearity and discrete nature, and enumerative techniques are often required to find the exact optimal solution. It is not difficult to show that IIP is NP-hard: consider the 0-1 knapsack problem $\max_{\mathbf{x}} \mathbf{v} \cdot \mathbf{x}$ among n_K objects such that $\mathbf{w} \cdot \mathbf{x} \leq 1$ and $\mathbf{x} \in \{0, 1\}^{n_K}$. Construct a graph $G_K = (N_K, A_K)$ with $N_K = \{1, 2, \ldots, n_K, n_K + 1\} \cup \{1', 2', \ldots, n'_K\}$ and $A_K = \{(1, 2), (2, 3), \ldots, (n_K, n_K + 1)\} \cup \{(1, 1'), (2, 2'), \ldots, (n_K, n'_K)\} \cup$

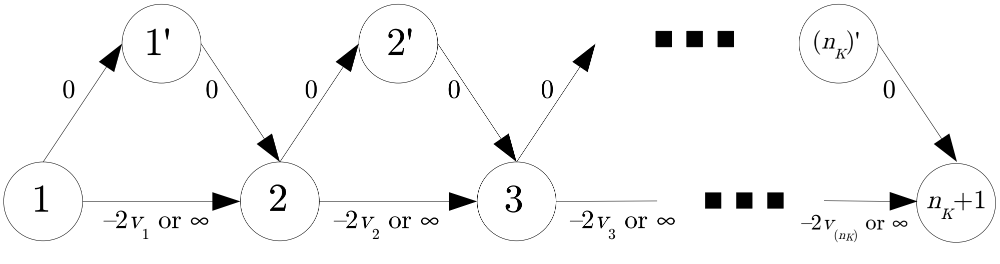


Figure 5.5: Reduction from the 0-1 knapsack problem.

$\{(1', 2), (2', 3), \ldots, (n'_K, n_K + 1)\}$ (see Figure 5.5). Each arc of the form $(i, i + 1) \in \{(1, 2), (2, 3), \ldots, (n_K, n_K + 1)\}$ exists in one of two states with equal probability; these states have cost $-2v_i$ and ∞, respectively. All other arcs have cost zero deterministically, and define $C_i = w_i$ for each node, along with $B = 1$. Consider solving IIP on G_K: if a node $i \in N_K$ is an information node, the optimal policy is clearly to follow $(i, i+1)$ if that arc has cost $-2v_i$, and to follow (i, i') otherwise. For non-information nodes j, the optimal policy is to always follow (j, j'), and the expected cost of any such policy is the negative of the knapsack objective when the objects corresponding to information nodes are selected. As this knapsack problem is well-known to be NP-hard, IIP must be NP-hard as well. Furthermore, as IIP is a special case of UIP and CIP, the NP-hardness of these problems follows immediately.

Thus no efficient, exact solution algorithms can be provided for these problems at present. Still, one way to determine the optimal set R^* is to simply calculate the total travel time resulting from each set in $\overline{R}$ being chosen as information nodes, and identifying the best such set. This is clearly inefficient, but network contraction makes enumeration computationally feasible for solving IIP or UIP on small- to medium-sized networks. That is, the contracted graph corresponding to each feasible set of information nodes is constructed, TD-OSP applied for each destination,

If $|R| \leq R_{max}$ for all feasible information sets R, then $O(n^{R_{max}})$ sets must be examined. Although this growth is polynomial in network size (assuming fixed R_{max}), a large planning network (such as those used to model Chicago, IL or Philadelphia, PA) can easily include over 10,000 nodes, and locating even three information nodes via enumeration would require more than a trillion iterations of network contraction and TD-OSP. Thus, it is still necessary to develop heuristic solution procedures.

Many heuristics employ the notion of a neighborhood to specify which feasible solutions are considered "adjacent" in a search procedure. In this section, the neighborhood $\mathcal{N}(R)$ of a set of information nodes R is defined as the set of feasible information node sets which differ from R by exactly one node. Returning to the example in Figure 5.3, where $R = \{5, 9\}$, $\mathcal{N}(R)$ is the union of the sets, $\{5\}$, $\{9\}$, $\{(5, i) : i \in N - \{5, 9\}\}$, $\{(i, 9) : i \in N - \{5, 9\}\}$, and $\{(5, 9, i) : i \in N - \{5, 9\}\}$, intersected with $\overline{R}$. In general this set is of size $O(nR_{max})$.

This suggests a local search heuristic, where one starts with an initial feasible set of information nodes, and considers each neighboring set. If any of them has a lesser objective function value, the least of these is chosen as the new incumbent solution, and the search repeated with the new neighborhood. If none has a lower objective function value, the current incumbent is declared a local optimum and the search halted. An initial feasible solution must be generated in some way; three approaches considered here are:

1. Labels representing the expected disutility from each origin to each destination are calculated for the "full-information" and "no information" cases (that is, where adaptive routing is allowed at each node, and where drivers must choose their route *a priori* using expected delays). The difference between these is defined as the *benefit* of information at node for that origin-destination pair; the total benefit is calculated by multiplying the benefit to each OD pair by its demand value, and summing. Proceeding in a greedy manner, construct the initial set R by repeatedly adding the nodes with the highest benefit-cost ratio, until doing so is no longer feasible.
2. Instead of choosing all of the nodes with highest benefit-cost ratio at the same time, proceed iteratively: after selecting the node i with the highest total benefit, re-calculate the "no information" labels by allowing information at node i along with the updated benefits, select the node with the highest total benefit which can feasibly be added to the initial set, and so on.
3. A purely random selection of nodes can be made for the initial solution.

Local search with this neighborhood definition is not guaranteed to find the optimal solution to IIP, as shown by the network in Figure 5.6. The only nodes where information can provide any benefit to travelers are nodes 2, 3, 5, and 6. By inspection, the optimal set is $R = \{2, 3\}$, with an expected travel time of 48. However, if the incumbent set is $R = \{5, 6\}$ (as would occur as the initial set under the first two decision rules), none of the neighboring information sets ($\{2, 5\}$,$\{2, 6\}$,$\{3, 5\}$,$\{3, 6\}$) reduce the expected travel time below its current value of 51. Thus, with this neighborhood definition, there can exist sets which are locally optimal, but not globally so.

Finally, one can apply a purely greedy approach: consider all feasible information sets of size one, and select the set R_1 providing the greatest reduction in the objective function per unit of cost, relative to the case where $R = \emptyset$. Next, consider all nodes which can be feasibly be added to R_1, and choose the set R_2 with providing the greatest objective function per unit of cost, relative to R_1. This procedure is repeated until no additional information nodes can feasibly be added. Thus, each iteration involves examining $O(n)$ new solutions. As with local search, this procedure need not produce the optimal solution, even when all nodes have equal cost, as seen by the network in Figure 5.6. If information can only be provided at one node, the best location is node 6, reducing the optimal expected travel time from 96 to 52. Given that information is provided at node 6, the best node to choose second is node 5, reducing expected travel time to 51; however, the optimal set of size two is $\{2, 3\}$, with expected delay 48.

From a practical standpoint, significant gains in computation time can often be obtained by judicious choice of the feasible sets $\overline{R}$, as influenced

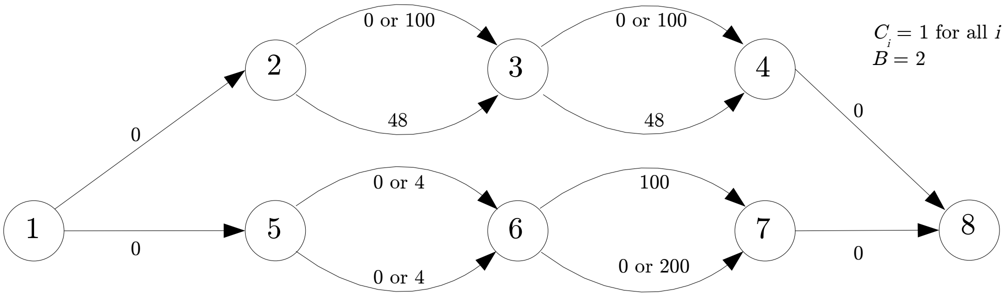


Figure 5.6: Network demonstrating how local search and the greedy heuristics can fail.

through the node costs C_i. For instance, there is no benefit to providing information at a node with only one exiting arc, because drivers at this node must choose the same arc regardless of any information received, and because any such information is only valid locally. Such nodes commonly exist where freeway onramps merge, and at certain intersections involving one-way streets or turn restrictions. In the Chicago Regional network, roughly five percent of the nodes can be excluded by this criterion. Since the number of feasible sets can grow exponentially with respect to the network size this saving can be significant: if $\overline{R} = N^3$, for instance, a time savings of nearly fifteen percent can be seen in an enumerative search. Separately, one may also be able to restrict attention *a priori* to a small subset of nodes, such as those adjoining freeway arcs and major arterials, leading to an even greater reduction in the size of the feasible set.

One should also note that all of these methods are highly parallelizable, which will decrease computation times substantially if available.

5.2.4 Demonstration

These algorithms were tested on the same test networks used in Chapter 4, whose characteristics can be seen in Table 4.1. While IIP can be studied on all four networks, memory and time considerations preclude analyzing UIP on the Chicago Regional network or analyzing CIP on the Barcelona or Chicago Regional networks. Generating contracted networks efficiently requires an all-pairs shortest path calculation to be made. This was accomplished using the Floyd-Warshall algorithm (Floyd, 1962), which required a negligible amount of computation time on the Sioux Falls network (less than 0.005 seconds), 1.14 seconds on the Anaheim network, 17.3 seconds on the Barcelona network, and 9.58 hours on the Chicago Regional network. Being common to all of the numerical tests that follow, these times are excluded from the run times reported for each solution method.

Each of the solution methods described in the previous section is implemented and tested. Local search is applied using each of the three rules for generating an initial information set; when the initial configuration is random, the search is repeated five times and the best solution chosen. Additionally, for comparison with a standard metaheuristic, simulated annealing is used to generate an information set, using the same neighborhood definition as the local search. The cooling schedule and other parameters are determined separately for each test network, adapting the procedure in Chiang and Russell (1996) to ensure that the initial probability of accepting a disimproving move is five percent, and that the number of iterations between cooling is equal to half of the neighborhood size. As before, computation times are reported for a 3.4 GHz Pentium 4 machine using Windows XP with 2 GB RAM. Furthermore, all algorithms are terminated after one hour of running time.

For each test case, the cost of providing information at each node is one cost unit, and the cases $B = 2$ and $B = 3$ are considered. That is, two feasible sets $\overline{R}$ are considered: R^2 (all sets of two information nodes) and R^3 (all sets of three information nodes). For IIP and UIP, arc delays are assumed to equal the free-flow travel time with probability 0.9, and three times the free-flow travel time with probability 0.1; for CIP, the free-flow travel times vary in the same manner, with the capacity constant; the well-known BPR relation is used to relate arc flows to travel times, with shape parameters $\alpha = 0.15$ and $\beta = 4$. Travel demand for UIP and CIP is the same as the standard network files; for IIP, the origin and destination are the two nodes farthest apart, in terms of shortest free-flow travel time.

Results from solving IIP on the four networks are shown in Table 5.4, showing the sets of information nodes found by the algorithms, the computation time needed to find these (in seconds), and the amount of benefits provided by information, relative to the benefits attainable by providing information everywhere. (That is, the difference between the expected travel delay with that information and the "no-information" expected travel delay, divided by the difference between the "full-information" and "no-information" expected travel delays.) The time required for finding the shortest path between each pair of nodes is reported in Table 4.1 and is not included in the computation times recorded here, in order to more clearly differentiate the impact of the algorithms which have a common initialization.

Several results are apparent. First and most notably, the greedy heuristic always found the best known solution, in substantially less time; this suggests that pitfalls such as those in Figure 5.6 are relatively rare in transportation networks, and that the sets of information nodes tend to "nest" in that optimal sets of one size are subsets of optimal sets of a larger size. On the other hand, the frequent failure of Local Search 3 (initialized randomly) to find the optimal information node sets, even with five restarts, suggests that local search quite often leads to non-globally optimal solutions if not initialized carefully.

Interestingly, the first two rules for determining the initial candidate set for a local search always produced identical sets of information nodes, and found the global optimum solutions, although rule one requires less computation time. This occurs because rule one only requires one application of TD-OSP on the whole network, while rule two requires one application per information node; while the benefits of iteratively updating disutility labels are not apparent in these networks. Unsurprisingly, enumeration quickly grows intractable; at the observed pace for the first hour of computation, identifying the optimal set of three information nodes on the Chicago Regional network would require more than a year.

Similar results are seen when solving UIP on the three smallest networks (Table 5.5). Note the substantial increase in computation time, since optimal

Table 5.4: Individual information provision (IIP) on test networks

(a) Sioux Falls

	$\lvert R\rvert = 2$			$\lvert R\rvert = 3$		
	Nodes	Time	Benefit	Nodes	Time	Benefit
Enumeration	3,12	0.02	56.5%	3,11,12	0.32	72.2%
Local Search 1	3,12	0.01	56.5%	3,11,12	0.02	72.2%
Local Search 2	3,12	0.01	56.5%	3,11,12	0.02	72.2%
Local Search 3	1,12	0.05	50.5%	1,11,12	0.10	69.3%
Greedy	3,12	0.00	56.5%	1,11,12	0.00	69.3%
Simulated Annealing	3,12	0.01	56.5%	3,11,12	0.32	72.2%

(b) Anaheim

	Nodes	Time	Benefit	Nodes	Time	Benefit
Enumeration	404,405	4.41	42.70%	305,404,405	957.93	56.0%
Local Search 1	404,405	0.35	42.7%	305,404,405	0.69	56.0%
Local Search 2	404,405	0.52	42.7%	305,404,405	0.84	56.0%
Local Search 3	180,404	0.28	22.9%	136,371,404	0.83	22.9%
Greedy	404,405	0.04	44.7%	305,404,405	0.21	56.0%
Simulated Annealing	404,405	0.20	42.7%	201,404,405	0.25	42.7%

(c) Barcelona

	Nodes	Time	Benefit	Nodes	Time	Benefit
Enumeration	249,1009	116	47.8%	1,351,783	3600*	47.8%
Local Search 1	249,1009	2.95	47.8%	249,306,1009	5.22	57.8%
Local Search 2	249,1009	5.23	47.8%	249,306,1009	9.62	57.8%
Local Search 3	550,1009	12.8	33.6%	909,921,1009	37.6	33.6%
Greedy	249,1009	0.27	47.8%	249,306,1009	0.57	57.8%
Simulated Annealing	963,1009	1.97	41.5%	249,826,1009	12.3	51.2%

*1.56% of feasible space explored in time limit.

(d) Chicago Regional

	Nodes	Time	Benefit	Nodes	Time	Benefit
Enumeration	2184,9883	3600*	18.0%	1,2184,9883	3600**	18.0%
Local Search 1	9446,9447	29.5	19.9%	2755,9476,12299	35.0	19.9%
Local Search 2	9446,9447	59.0	19.9%	2755,9476,12299	94.4	19.9%
Local Search 3	7051,9883	15.1	14.9%	1896,3358,9883	60.5	14.9%
Greedy	6826,9883	3.04	19.9%	6826,8625,9883	6.57	23.5%
Simulated Annealing	6822,9883	31.5	18.5%	2184,8625,9883	34.8	21.6%

*66.6% of feasible space explored in time limit

**0.01% of feasible space explored in time limit

policies must be found for each destination in the network, not just one. The comments which applied to IIP are mainly applicable here as well. Although the first two decision rules for initializing the local search seem to produce different results for locating three information nodes in the Barcelona network, this is an artifact introduced by the one-hour time limit and the greater time needed to initialize rule two. Given more time to proceed, Local Search 2 would have followed the same search trajectory as Local Search 1 in this network.

Table 5.6 shows the results from solving CIP on the Anaheim and Sioux Falls networks. Interestingly, for the Anaheim network, using a random seed for the local search yielded a better two-information node solution than was found by any of the other heuristics, the only time that this heuristic found a better solution than the others. Comparing UIP and CIP, one sees the benefits attainable from only two or three information nodes are higher in the Sioux Falls network when congestion effects are present, but lower in the Anaheim network. This may be due to differences in the congestion level on these networks: the average volume-to-capacity ratios for the no-information equilibrium assignment in these networks are 1.48 for Sioux Falls, and 0.32 for Anaheim. Again note the significant increase in computation time needed to solve this problem, as evaluating any feasible solution involves an equilibration, and no network contraction is available to speed the process.

In all cases, note that a sizable portion of the total possible benefits from information provision can be achieved even when only providing information at two or three nodes.

Table 5.5: Uncongested information provision (UIP) on test networks

(a) Sioux Falls

	$\lvert R\rvert = 2$			$\lvert R\rvert = 3$		
	Nodes	Time	Benefit	Nodes	Time	Benefit
Enumeration	10,16	1.41	29.0%	10,15,16	18.57	38.3%
Local Search 1	10,16	0.87	29.0%	10,15,16	3.31	38.3%
Local Search 2	10,16	0.97	29.0%	10,15,16	3.60	38.3%
Local Search 3	10,16	2.27	29.0%	10,15,16	5.71	38.3%
Greedy	10,16	0.20	29.0%	6,10,14	0.48	26.8%
Simulated Annealing	10,16	0.75	29.0%	10,15,16	3.66	38.3%

(b) Anaheim

	$\lvert R\rvert = 2$			$\lvert R\rvert = 3$		
	Nodes	Time	Benefit	Nodes	Time	Benefit
Enumeration	91,232	648.3	18.8%	2,91,232	3600*	25.6%
Local Search 1	91,232	20.8	18.8%	91,232,236	57.8	25.6%
Local Search 2	91,232	24.0	18.8%	91,232,236	65.7	25.6%
Local Search 3	227,232	55.9	12.5%	91,95,232	129.5	19.2%
Greedy	91,232	4.4	18.8%	91,232,236	8.1	25.6%
Simulated Annealing	91,232	9.2	18.8%	91,232,236	27.6	25.6%

*3.97% of feasible space explored in time limit.

(c) Barcelona

	$\lvert R\rvert = 2$			$\lvert R\rvert = 3$		
	Nodes	Time	Benefit	Nodes	Time	Benefit
Enumeration	1,766	3600*	5.06%	1,8,766	3600**	5.14%
Local Search 1	555,766	1733	11.7%	555,673,766	3600	20.6%
Local Search 2	555,766	2425	11.7%	72,766,887	3600	11.8%
Local Search 3	306,766	3000	6.32%	210,366,762	3600	8.86%
Greedy	555,766	303	11.7%	555,673,766	623	20.6%
Simulated Annealing	682,762	888	9.24%	555,676,816	1610	15.3%

*3.90% of feasible space explored in time limit.
**0.01% of feasible space explored in time limit

Table 5.6: Congested information provision (CIP) on test networks

(a) Sioux Falls

	$\lvert R\rvert = 2$			$\lvert R\rvert = 3$		
	Nodes	Time	Benefit	Nodes	Time	Benefit
Enumeration	10,15	40.4	35.4%	10,11,15	310	45.1%
Local Search 1	10,15	19.0	35.4%	10,11,15	55.8	45.1%
Local Search 2	10,15	13.0	35.4%	10,11,15	31.6	45.1%
Local Search 3	10,15	64.3	35.4%	10,15,24	105	43.3%
Greedy	10,15	6.0	35.4%	10,11,15	9.5	45.1%
Simulated Annealing	10,15	23.8	35.4%	10,11,15	47.0	45.1%

(b) Anaheim

	$\lvert R\rvert = 2$			$\lvert R\rvert = 3$		
	Nodes	Time	Benefit	Nodes	Time	Benefit
Enumeration	21,319	3600*	11.0%	1,21,319	3600**	11.0%
Local Search 1	319,355	463	13.6%	319,355,407	1030	17.0%
Local Search 2	319,355	467	13.6%	319,355,407	1033	17.0%
Local Search 3	388,389	1331	15.8%	86,302,319	2403	11.0%
Greedy	319,355	147	13.6%	319,355,407	236	17.0%
Simulated Annealing	319,355	169	13.6%	268,319,355	554	11.0%

*20.9% of feasible space explored in time limit.

**0.15% of feasible space explored in time limit

5.2.5 Conclusion

This section addressed the problem of choosing the optimal locations to provide real-time traffic information, in three different forms: routing of an individual vehicle, routing of multiple vehicles in an uncongested system, and multiple-vehicle equilibrium in a congested network. As even the simplest of these problems is NP-hard, heuristics were developed to solve each of these problems. For the two simplest cases, a network contraction procedure allows rapid evaluation of candidate solutions. These heuristics were then tested in networks of varying sizes, showing that a substantial portion of the benefits of information are available even when providing information only at two or three nodes.

Importantly, studying this improvement strategy from this perspective is made possible by the NL-OSP and NL-UER algorithms, which allow potential information node configurations to be examined consistently, allowing for anticipation of this information in advance.

5.3 Congestion Pricing

In recent years, pricing of highway driving (tolling) has attracted much political and institutional attention for a variety of reasons, including its potential as an alternate revenue stream, the introduction of technologies allowing efficient toll collection and dynamic pricing, and consideration of public-private partnerships. To support the process of determining appropriate prices, a large amount of research has been conducted to provide guidance on how users respond to prices, and how they should be set to achieve particular objectives. From the standpoint of maximizing social welfare, the fundamental notion, originated by Pigou (1920), is that economic efficiency occurs when the cost faced by each traveler equals the marginal social cost of his or her trip.

Traditionally, this marginal cost is determined by assuming a separable and differentiable volume-delay function (VDF) mapping travel demand to travel delay on each roadway segment, a homogeneous population of user-optimizing travelers with the same value of time, and commonly-known network structure and travel demand. Within this framework, the marginal cost of tripmaking is readily calculated, along with the associated Pigouvian tolls.

However, the concept of uncertain roadway supply has not yet been integrated into pricing models intended for use in large-scale networks. Further, given recent technological advances, network operators and planners may wonder whether prices should be dynamically varied in response to traffic incidents or other disruptions, and, if so, how this variation should occur. For instance, one might argue that tolls on a facility should increase if an incident occurs, to discourage additional vehicles from entering and exacerbating the resulting congestion. However, in response, one might argue that users paying a higher toll should expect a higher level of service, as this is one of the usual arguments provided to gain public support for congestion pricing. Wouldn't travelers resent paying a higher toll, while most likely still experiencing greater-than-average delay? Or is there some way to account for uncertainty on a daily basis without varying tolls?

As with most problems concerning uncertainty, the question of information is key: who knows what, when they make their decisions? For instance, the issue of resentment for higher tolls during an incident is greatly decreased if operators can communicate to motorists the presence and severity of the incident. This research presents four possible scenarios relating to the information available to motorists when choosing a travel route, and to the ability of the network manager to adjust the toll in response to network conditions.

The key contribution here is the development of pricing methods to apply in the presence of operational supply uncertainty and risk-averse travelers, for several information provision scenarios. The remainder of this section is organized as follows: Section 5.3.1 discusses prior literature related to pricing, travel time uncertainty, and user attitudes to risk. Section 5.3.2 describes the modeling approach, introducing appropriate notation and defining the four information scenarios. Section 5.3.3 presents solution methods for each of these scenarios. The models described thus far make a number of simplifications, and Section 5.3.4 discuss how they can be adapted to account for correlated arc states, user heterogeneity, and elastic demand. Section 5.3.5 demonstrates the basic model using the well-known Sioux Falls test network, and suggests that constant tolls should not be used to address nonrecurring congestion.

5.3.1 Literature Review

This section summarizes prior related work, focusing on three areas: pricing under uncertain network conditions; the impact of reliability on route choice; and how pricing and reliability interact. In this light, the contribution of this section should be more apparent, and is briefly discussed at the end of the section.

The question of how to appropriately price freeway facilities in uncertain environments is still very open. Yang (1999a) considered the problem of determining optimal prices when users behave according to the stochastic user equilibrium principle, where uncertainty lies in user perception, rather than system conditions. It is known that there need not exist a set of tolls that can drive a stochastic user equilibrium traffic flow pattern to a system optimal one (Akamatsu and Kuwahara, 1988; Smith et al., 1994). Yang (1999b) also considered how road pricing can be combined with advanced traveler information systems which inform users of system conditions. A number of numerical experiments were performed in a small test network, from which the author concluded that the two technologies "complement each other and that their joint implementation can reduce travel time more efficiently." Separately, de Palma and Lindsey (1998) considered information provision under three different scenarios: free access, non-responsive congestion pricing, and dynamic pricing based on congestion levels. These authors explicitly considered capacity uncertainty in all of their models, but in a simplified setting without network effects and multiple origins and destinations. Under these assumptions, when pricing is dynamic and responsive to congestion, these authors showed that better information always improves welfare. A key result of Mohring and Harwitz (1962) is that marginal-cost pricing generates enough revenue to provide socially-optimal facility capacity; Lindsey (2008) showed that this result generalizes to the case of uncertain capacity if drivers are perfectly informed and tolls are responsive, or under imperfect information if tolls are set according to the same information drivers have, and if the price elasticity of demand does not vary with system conditions.

It is clear that reliability plays a significant role in route choice decisions; however, there is no consensus on how “reliability” should be defined. Usually, this is done in relation to the distribution of possible path costs. For instance, Small et al. (2005) and Liu et al. (2007) used the difference between the 80th- and 50th-percentile travel times, while Pinjari and Bhat (2006) used the maximum additional time that could be needed, compared to a typical case. Gao (2005), on the other hand, assumed a piecewise-linear utility function to model risk aversion. de Palma and Picard (2005) considered four utility function specifications to represent risk aversion: penalizing the standard deviation of travel time, penalizing travel time variance, constant relative risk aversion, and constant absolute risk aversion. Bates et al. (2001) and Noland and Polak (2002) provided overviews of theoretical and empirical research in travelers’ valuations of travel time reliability. Typically, travelers’ sensitivity to reliability is comparable to their sensitivity to increased travel time; for instance, Small et al. (2005) estimated a \$21.46/hr value of time, and a \$19.56/hr value of reliability, using data from SR-91 in California.

In contrast to the utility-based methods above, Avineri and Prashker (2003) accounted for uncertainty in route choice using cumulative prospect theory, the Fudenberg-Levine learning model (Fudenberg and Levine, 1998), a behavioral "reinforcement learning" model, and a novel cumulative prospect theory learning model. Chan and Lam (2005) took a completely different approach, using a novel concept of user equilibrium based on "path preference indices."

Several researchers have studied the interaction between pricing and facility reliability. The research in this area has been descriptive (attempting to evaluate how pricing affects facility reliability, or studying pricing to discern how travelers value reliability) rather than prescriptive (how should prices be set to maximize reliability or traveler welfare). Supernak et al. (2003) performed a before-after study of the I-15 FasTrak value pricing project in San Diego, California, looking specifically at changes in travel time and travel reliability, measured as the 99th-percentile of travel time. Using this definition, they found substantial improvements in reliability after implementation. Liu et al. (2004) used freeway loop data from California State Route 91 to estimate a random-parameters logit model with two alternatives, free and tolled lanes. Travel time, reliability (defined as the difference between the 80th- and 50th-percentile travel times, approximately one standard deviation in several common probability distributions.), and toll amount are used as alternative-specific variables. They applied a genetic algorithm to estimate the logit parameters, resulting in an estimated value of time of \$13/hour and an estimated value of reliability of \$21/hour. Brownstone and Small (2005) also considered the I-15 and SR-91 project, and used both stated and revealed preference data. They estimated a relatively high value of time (between \$20/hour and \$40/hour for the morning commute) based on revealed preference data, and a much lower value (around \$12/hour) from stated preference surveys.

Although considerable research exists on pricing and network routing under uncertainty, relatively little research combines the two, especially regarding travelers' risk attitudes and/or valuation of reliable travel. Further, that which has been done has typically involved simplified settings and small networks, which admit important analytical results but which are less useful in guiding implementation of pricing policies. Emmerink et al. (1996) showed that no subsidy towards information provision is needed to maximize social welfare, given first-best congestion pricing and costly information provision. Verhoef et al. (1996) simulated a two-arc network under different pricing and information scenarios, concluding that information provision and "flat" (unresponsive) tolls are nearly as effective as perfectly responsive tolls. Kobayashi and Do (2005) considered a simple network with non-overlapping routes and a single origin-destination pair, and showed that perfect information and *ex post* tolls maximize social welfare.

In this light, this section's primary contribution is the development of models to identify optimal tolls in large-scale networks, when roadway capacity is stochastic. In these models, corresponding to different information scenarios, route choice is endogenous (i.e., traffic assignment and equilibrium are included), and prices inducing system-optimal (or approximately system-optimal) arc flows are sought.

5.3.2 Problem Statement

This section mathematically describes the pricing model assumptions, along with four information scenarios. Section 5.3.2.2 addresses the issue of how to model users' valuation of reliability in route choice, and reviews the equilibrium concepts that will be applied to determine user response to a set of tolls. Section 5.3.2.3 describes the network manager's goal, and section 5.3.2.1 defines the information scenarios we analyze. An assumption of user homogeneity is taken at the start to emphasize the key points of the basic model; the implications of relaxing this is discussed in Section 5.3.4.

Generally, pricing problems are a type of Stackelberg game, in which a regulator acts as a "leader" by establishing a set of tolls, to which individual drivers ("followers") respond by choosing preferred routes. We adopt the same perspective, but with the additional complication of uncertain network conditions. As described in Section 5.3.2.2, a generalized cost function is assumed for travelers, accounting for average travel time, reliability, and monetary tolls. The goal of the network manager is to choose tolls so as to bring the user equilibrium and system optimal arc flows into alignment (i.e., incorporating externalities into individual costs).

Depending on the information provision scenario, the tolls and arc flows may vary according to the network realization ω; when needed, a superscript will denote which realization a toll or flow value corresponds to.

5.3.2.1 Information Scenarios

The question of which agents have access to what information plays a defining role in determining the structure and results in a stochastic optimization model. In this problem, there are two types of agent: the network manager, who establishes the tolls, and the users, who choose routes.

Initially, we consider two information scenarios for each agent (leading to four scenarios in total): a “no information” case, in which the agent is unaware of the network realization before making the decision, and a “fully informed” case, in which the agent learns the exact network realization. This information is assumed to be perfectly accurate and fully trusted by all agents. Notationally, these scenarios are distinguished by the presence or absence of a subscript or superscript corresponding to the arc state s: decision variables which vary by realization are marked as such, while decision variables which do not vary by realization do not have such an indication.

For the manager, the “no information” case is identical to one in which tolls cannot vary between states; for this reason, the manager’s information is denoted as either RT (responsive tolls, for full information) or UT (unresponsive tolls, for no information or when responsive tolling is impossible). For the users, the “no information” case implies that the route must be chosen before learning the network realization and the tolls; the “full information” case implies that the tolls and network realization are learned upon reaching the upstream node of the arc. These user information scenarios are denoted NI and AR (adaptive routing), respectively. Thus, the four information scenarios here are AR/UT, AR/RT, NI/UT, and NI/RT (indicating the users' information first and the regulator's information second):

Adaptive Routing/Unresponsive Tolls This scenario represents a case in which users obtain full (lcoal) information at each node, but tolls cannot vary in response to their choices or the network realization. This can occur either because the regulator is unaware of network conditions, or because the regulator is not allowed to change the tolls. More precisely, users learn the prevailing travel time functions and their relevant parameters (such as capacity), and the equilibrium state arising from this common knowledge is sought. Thus, the flows $\mathbf{x_{ij}^s}$ vary according to the arc state, but the tolls $\boldsymbol{\tau}$ do not.

Adaptive Routing/Responsive Tolls This scenario represents maximum information for all decision makers in the problem: travelers learn each arc's state upon arriving at its upstream node, and the network manager can set different tolls for different states. This is similar to the AR/UT case, except that the tolls $\boldsymbol{\tau^s}$ are also allowed to vary according to the arc state.

No Information/Unresponsive Tolls In this information scenario, neither users nor the manager can vary their decisions according to the network realization. In this case, although the tolls are known and fixed from day to day, users are unaware of the network realization when making the routing decision, and thus have no reason to vary their decision from day to day.

No Information/Responsive Tolls In this scenario, users are unaware of the network realization, even though the network manager can vary the tolls responsively. However, varying tolls cannot provide any additional benefit to users if they do not learn of them before they choose a route. Essentially, this scenario is identical to NI/UT, since the manager cannot induce a superior flow pattern by varying tolls if users cannot respond in turn.

5.3.2.2 User Behavior

All travelers are assumed to be homogeneous, and value travel on any path p according to a generalized cost function C depending on arc travel times and tolls. When travelers know arc states before departing, their cost function is simply the weighted sum of travel time and toll paid: This clearly separates by arc, leading to arc cost functions

$$c_{ij}^s = VOTT \times t_{ij}^s + \tau_{ij}^s \tag{5.2}$$

where $VOTT$ represents the value of travel time.[3]

On the other hand, if arc states are unknown to travelers when they

---

[3] The attentive reader will notice a change in notation made in this section (and only for this section), where c_{ij} now represents the generalized cost on a arc, and χ_{ij} the capacity. This change is adopted to reflect the customary notation for pricing problems, and because the notation for arc capacity is only sparingly used in this section.

choose routes, the cost function is

$$c_{ij} = VOATT \times E_s[t_{ij}^s(x_{ij}^s)] + VOTR \times Var_s[t_{ij}^s(x_{ij}^s)] + \tau_{ij} \tag{5.3}$$

where $VOATT$ and $VOTR$ represent the value of average travel time (not necessarily the same as $VOTT$), and the value of travel reliability, respectively.

With the cost function defined, a more formal demonstration is given that variable tolls cannot provide additional benefit if users are uninformed: if travelers are unaware of tolls when choosing a route, equation (5.3) must be modified, e.g., $c_{ij} = -VOATT \times E[\tilde{t}_{ij}] - VOTR \times Var[\tilde{t}_{ij}] - g(\tilde{\tau}_{ij}, p_{ij}^s)$ for some function g (including, for instance, the expected value and variance of the tolls $\tilde{\tau}$ which are now perceived as random variables), and let $\mathbf{x}(\boldsymbol{\tau})$ represent the equilibrium arc flow vector obtained for tolls $\boldsymbol{\tau}$ over all network realizations. The same flow vector can be replicated under the NI/UT scenario by defining a random variable $\tilde{\upsilon}$ which takes the values τ_{ij}^s with probabilities p_{ij}^s, and setting tolls $\tau_{ij} = g(\tilde{\upsilon}, p_{ij}^s)$. The generalized cost on each arc is the same under this construction as in the NI/RT scenario, and thus $\mathbf{x}$ remains an equilibrium.

Note that in the case of information provision, cost function (5.2) is a special case of (5.3), in which the variance in arc cost vanishes. Thus, these two utility functions are consistent, and meaningful comparisons can be made between tolls derived for both scenarios.

This specification only considers within-day travel time uncertainty: even if travel times vary widely from day to day, equation (5.2) is used as long as travelers learn the exact realization before departing. That is to say, the inherent value of stable travel times over multiple days, such as the establishment of routine habits, is excluded from consideration. Rather, the focus in this section is the cost of imperfect knowledge of travel times on a given day, leading to earlier departure times (leaving a "safety margin") or running the risk of late arrival.

As described in Section 5.3.1, some researchers prefer to use standard deviation instead of variance, since it has common units with expected travel time. In this section, we opt to use variance for three reasons: first, mean-variance models are commonly used to model risk in domains such as finance (see, for instance, Markowitz, 1952; Sternbach, 2001); second, it is more convenient mathematically, since variances add linearly under the independence assumption (i.e., $Var[A+B] = Var[A]+Var[B]$), allowing ready computation of path variance; finally, since variance is the square of standard deviation, this model places increasingly greater weight on travel reliability as travel times become more uncertain.

As a side note, one may not need to explicitly sum over all arc states when evaluating $E_s[t_{ij}^s]$ and $Var_s[t_{ij}^s]$, depending on the cost function. For instance, assuming a standard Bureau of Public Roads cost function of the form $t_{ij}^s = t_{ij}^0(1 + \alpha(x_{ij}/\chi_{ij}^s)^\beta)$ in which only the capacity parameter χ_{ij}^s varies by realization, it is readily verified that $E_s[t_{ij}^s] = t_{ij}^0(1+\alpha\phi_{ij}x_{ij}^\beta)$ and $Var_s[t_{ij}^s] = \theta_{ij}(\alpha t_{ij}^0 x_{ij}^\beta)^2$, where $\phi_{ij} = \sum_{s\in S_{ij}} (\chi_{ij}^s)^{-\beta} p(s_{ij})$ and $\theta_{ij} = \sum_{s\in S_{ij}} (\chi_{ij}^s)^{-2\beta} p(s_{ij}) - \phi_{ij}^2$ are arc-specific constants that need only be calculated once, independent of demand values and route choices.

5.3.2.3 Network Manager Behavior

In our model, the network manager's goal is to minimize the total *travel time*-related costs experienced by travelers; that is, when users learn arc states before departing, the network manager seeks to minimize

$$\sum_{(i,j)\in A} x_{ij}(VOTT \times E_s[t_{ij}^s]) \tag{5.4}$$

In contrast, when users are ignorant of arc states when choosing routes, the network manager minimizes

$$\sum_{(i,j)\in A} x_{ij}(VOATT \times E_s[t_{ij}^s] + VOTR \times Var_s[t_{ij}^s]) \tag{5.5}$$

This is done by setting the tolls $\boldsymbol{\tau}$ in such a manner as to bring the user and system objectives into alignment. Note that the network manager's goal does not include minimizing the toll-related costs. This assumes that toll revenues are effectively returned to the region in which they are collected with minimal administrative burden, perhaps through additional infrastructure spending or reduced taxation.

5.3.3 Solution Methods

This section describes methods for finding tolls that bring the user and system objectives into alignment, for the three information scenarios AR/UT, AR/RT, and NI/UT. (As demonstrated in the previous section, NI/RT is a special case of NI/UT, and need not be considered separately). All of these are based on the Pigouvian principle that externalities should be incorporated into user costs or, equivalently, that average cost equal the marginal social cost. Since the marginal cost of travel on a arc (i, j) is $d(x_{ij}c_{ij})/dx_{ij} = c_{ij} + x_{ij}(dc_{ij}/dx_{ij})$, the Pigouvian toll is $x_{ij}(dc_{ij}/dx_{ij})$ with the cost function appropriate to the information scenario, and system optimal flows $\mathbf{x}$.

5.3.3.1 Adaptive Routing/Unresponsive Tolls

The AR/UT scenario is the most complicated to solve, for several reasons. First, the constraint that the tolls must be the same for all network realizations prevents a simple decomposition by network realization. Second, the requirement that the flows for every network realization be in equilibrium imposes nonconvexity on the toll-setting problem (see, for instance, Labbé et al., 1998), a problem confounded by the nonlinearity of travel time functions, which makes it unlikely that a globally optimal solution can be found. Third, since the flows vary according to the network realization, and since the number of network realizations is very large, computation of the objective function for even a single set of tolls is nontrivial. Essentially, the desired toll vector solves the program

$$\min_{\boldsymbol{\tau}=[\tau_{ij}]} \sum_{(i,j)\in A} \sum_{s\in S_{ij}} t^s_{ij}(x^s_{ij})x^s_{ij} \tag{5.6}$$

$$\text{s.t.} \quad \tau_{ij} \geq 0 \qquad \forall (i,j) \in A \tag{5.7}$$

$$\mathbf{x} \in Eq(VOTT \times \mathbf{t} + \boldsymbol{\tau}) \qquad \forall s \in s \tag{5.8}$$

where $Eq(VOTT \times \mathbf{t} + \boldsymbol{\tau})$ represents the set of UER arc usages for delay functions $\mathbf{t^s}$ and tolls $\boldsymbol{\tau}$. This is a nonlinear mathematical program with equilibrium constraints (MPEC), which is known to be difficult to solve, for the reasons mentioned above. For this reason, approximately optimal tolls are sought. One option is to use a generic metaheuristic, such as simulated annealing or tabu search. Another choice is to use problem-specific heuristics, two of which are described below.

Heuristic 1 (H1) is to use simple averaging: "network states" are sampled by randomly selecting a state for each arc, finding first-best marginal-cost tolls for the resulting deterministic problem, and obtaining the final state-dependent toll vector by averaging the first-best tols prevailing in each state.

Heuristic 2 (H2) is somewhat more involved, and allows tolls to vary by realization, penalizing this variation in the objective function with a positive constant M:

$$\min_{\boldsymbol{\tau}=[\tau_{ij}^s]} \sum_{(i,j)\in A} \sum_{s\in S_{ij}} \left[t_{ij}^s(x_{ij}^s)x_{ij}^s + M(\tau_{ij}^s - \overline{\tau}_{ij})^2\right] \tag{5.9}$$

$$\text{s.t.} \quad \tau_{ij}^s \geq 0 \qquad \forall (i,j) \in A, s \in S_{ij} \tag{5.10}$$

$$\mathbf{x} \in Eq(VOTT \times \mathbf{t} + \boldsymbol{\tau}) \tag{5.11}$$

where $\overline{\tau}_{ij}$ is the average toll on arc (i,j) across all states. Linearizing the objective function, we seek a vector of tolls satisfying

$$\frac{d\left(\sum_{(k,\ell)\in A} x_{k\ell}^s t_{k\ell}^s(x_{k\ell}^s)\right)}{d\tau_{ij}^s} = -2M(\tau_{ij}^s - \overline{\tau}_{ij}) \tag{5.12}$$

for all $(i, j) \in A$. Applying the chain rule to the left-hand side, an analytical solution for each τ_{ij}^s is obtained:

$$\tau_{ij}^s = \overline{\tau}_{ij} - \frac{1}{2M} \sum_{(k,\ell)\in A} \frac{dx_{k\ell}}{d\tau_{ij}} \left[t_{k\ell}^s(x_{k\ell}^s) + x_{k\ell}^s \frac{dt_{k\ell}^s}{dx_{k\ell}^s} \right] \tag{5.13}$$

where the $dx_{ij}^s / d\tau_{k\ell}^s$'s are estimated by perturbing the toll on each arc slightly and observing the resulting change in equilibrium arc flows. This suggests an iterative procedure in which subproblems are successively solved, with increasing values of the penalty constant M (clearly, as $M \to \infty$ the realization-specific tolls converge to a common value, which is returned as the solution). This is shown formally in Algorithm 6, where M_0 and ϵ respectively denote the initial value of the penalty constant, and the convergence criterion.

Algorithm 6 Heuristic 2 for AR/UT

```
{Initialization}
for all $(i, j) \in A, s \in S_{ij}$ do
    $\tau_{ij}^s \leftarrow$ first-best marginal cost state-dependent toll
end for
$\overline{\tau}_{ij} \leftarrow \sum_{s \in S_{ij}} p_{ij}^s \tau_{ij}^s$
$M \leftarrow M_0$
{Iteration}
while $\max_{(i,j),s} |\tau_{ij} s - \overline{\tau}_{ij}| > \epsilon$ do
    for all $(i, j) \in A, s \in S_{ij}$ do
        $\tau_{ij} \leftarrow \overline{\tau}_{ij} - \frac{1}{2M} \sum_{(k,\ell)\in A} \frac{dx_{k\ell}}{d\tau_{ij}} \left[ t_{k\ell}^s(x_{k\ell}^s) + x_{k\ell}^s \frac{dt_{k\ell}^s}{dx_{k\ell}^s} \right]$
    end for
    $M \leftarrow 2M$
end while
return $\boldsymbol{\overline{\tau}}$
```

5.3.3.2 Adaptive Routing/Responsive Tolls

In the AR/RT scenario, tolls are allowed to vary for different arc states, greatly simplifying the problem at hand. As shown in Unnikrishnan (2008), the system-optimal recourse arc usages can be stated as the solution to the optimization problem

$$\min \sum_{(i,j)\in A} \sum_{s\in S_{ij}} x^s_{ij} c^s_{ij}(x^s_{ij}) \tag{5.14}$$

where the x^s_{ij} map from feasible policy flows. Lagrangianizing the demand constraint, and expressing the objective function in terms of the policy flows $\mathbf{y}$ (through the relation $\mathbf{x} = A\mathbf{y}$ for the incidence matrix defined in Section 3.2.4), the Karush Kuhn-Tucker (KKT) conditions for this program include

$$\sum_{(i,j)\in A} \sum_{s\in S_{ij}} a_{\pi,ijs} \left(c^s_{ij}(x^s_{ij}) + x^s_{ij} \frac{dc^s_{ij}}{dx^s_{ij}} \right) - \kappa_{uv} \geq 0 \quad \forall (u,v) \in D, \pi \in \Pi_{uv} \tag{5.15}$$

$$y_\pi \left[\sum_{(i,j)\in A} \sum_{s\in S_{ij}} a_{\pi,ijs} \left(c^s_{ij}(x^s_{ij}) + x^s_{ij} \frac{dc^s_{ij}}{dx^s_{ij}} \right) - \kappa_{uv} \right] = 0 \quad \forall (u,v) \in D, \pi \in \Pi_{uv} \tag{5.16}$$

substituting x^s_{ij} for $\sum_{(u,v) in D} \sum_{\pi\in\Pi_{uv}} a_{\pi,ijs} y_\pi$ when convenient.

This is strongly reminiscent of the KKT conditions for the user equilibrium with recourse problem, which include

$$\sum_{(i,j)\in A} \sum_{s\in S_{ij}} a_{\pi,ijs} c^s_{ij}(x^s_{ij}) - \kappa_{uv} \geq 0 \qquad \forall (u,v) \in D, \pi \in \Pi_{uv} \tag{5.17}$$

$$y_\pi \left[\sum_{(i,j)\in A} \sum_{s\in S_{ij}} a_{\pi,ijs} c^s_{ij}(x^s_{ij}) - \kappa_{uv} \right] = 0 \qquad \forall (u,v) \in D, \pi \in \Pi_{uv} \tag{5.18}$$

and where the κ_{uv} indicate the disutility on the least expected-cost policy connecting u to v. Note that the effect of the two constraints is to ensure that policies are only used if their disutility is minimal.

Thus, if the state-dependent tolls t^s_{ij} are set to $x^s_{ij}(dc^s_{ij})/dx^s_{ij})$, the user optimal and system optimal objectives coincide, demonstrating that

$$\tau^s_{ij} = x^s_{ij} \left. \frac{dc^s_{ij}}{dx^s_{ij}} \right|_{x^s_{ij}} \tag{5.19}$$

is the proper toll to set in this state.

5.3.3.3 No Information/Unresponsive Tolls

The NI/UT scenario is the simplest case, since only one vector of network flows and one vector of tolls is needed. Solving for system-optimal arc flows and marginal-cost prices using cost function (5.3) gives optimal tolls for this information scenario.

5.3.4 Heterogeneous Users

Users are not uniform in their valuation of travel time and reliability. Instead, one can imagine that the parameters $VOTT$, $VOATT$, and $VOTR$ can be represented as (possibly correlated) distributions over the population. The idea that these parameters vary in the population is both intuitive, and has been empirically demonstrated in multiple stated and revealed preference surveys. If the value of reliability is not considered (or if $VOTR$ is assumed to be an affine function of $VOATT$), the bicriterion equilibrium and pricing framework presented in Dial (1996, 1997, 1999a,b) suffices for identifying welfare-maximizing tolls. For the "full information" cases, where path travel time variance vanishes, nothing further is needed.

On the other hand, including reliability as a third criterion (alongside toll and average travel time, as in the "no information" case) introduces a few more complications. With a few suitable modifications, Dial's approach can be applied to the tricriterion case, as described in this section. The basic concept involves identifying a set of efficient paths and assigning trips accordingly, because only a small set of paths will be used by travelers regardless of their values of travel time and reliability. This assignment process is then applied iteratively to find a user equilibrium and welfare-maximizing tolls.

These assignments are performed using prevailing attributes. That is, arc tolls, travel time means, and travel time variances are temporarily assumed to be fixed and independent of traffic flow, and the efficient paths are those which are least-cost paths with respect to some values of $VOATT$ and $VOTR$. Each path π has an associated *tricriterion vector* $\mathbf{P}^\pi$, whose three components are the toll, mean travel time, and travel time variance on path π. Plotting all vectors $\mathbf{P}$ in the first octant, the efficient paths are seen to be the lower extreme points of their convex hull.

One can identify efficient paths using several techniques from multiobjective optimization, such as weighted objective functions (Geoffrion, 1968), the ϵ-constraining method (Haimes et al., 1971), and a decomposition method using the Chebyshev metric (Eswaran et al., 1989). Since our utility functions are simply linear functions of the path attributes, the weighted objective func-

Path	τ	$E[t]$	$Var[t]$
P1	2	0	1
P2	0	0	2
P3	2	1	0
P4	0	2	0

Table 5.7: Four paths used in tricriterion demonstration

tion method as adapted by Dial (1996) generates these paths efficiently for the bicriterion problem.

Dial (1996) speculates that the efficient frontier of the tricriterion problem is a "triangulated convex surface, with each vertex representing a path." This point requires careful definition, since the convex hull of the set of efficient paths may include faces adjoining more than three vertices. Consider a network consisting of four paths, whose tolls, mean travel times, and travel time variances are shown in Table 5.7. All of these paths are efficient, and their convex hull is a plane segment with four vertices, as seen in Figure 5.7. Of course, one can express this quadrilateral plane segment as the union of triangular plane segments (say, triangles P1-P2-P4 and P1-P3-P4), in which case the efficient frontier is trivially seen to consist of a convex union of triangles. This representation is useful algorithmically, and adopted throughout this section.

Once a set of efficient paths is identified, one must assign heterogeneous users to these paths. In the bicriterion case, travelers are partitioned according to their values of travel time, and Dial (1996) shows that each path can be associated with an interval in $\mathbb{R}_+$, and each user chooses the path associated

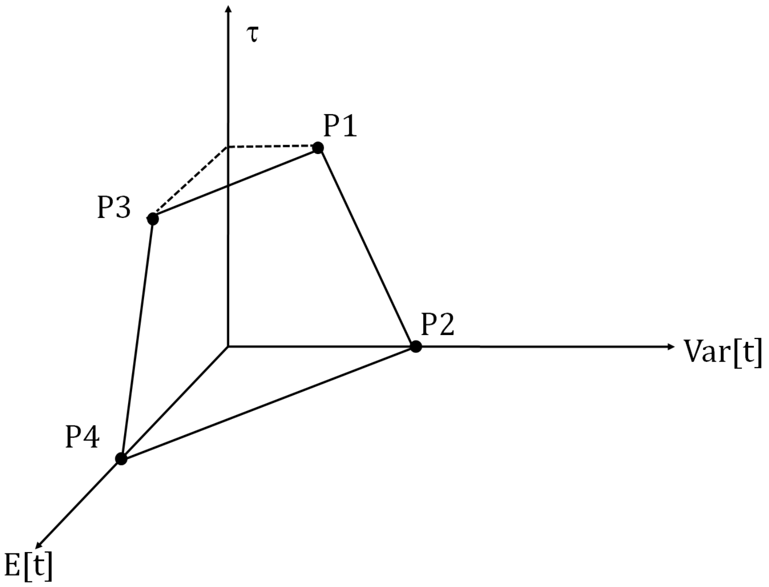


Figure 5.7: Four efficient paths and their convex hull

with the interval containing their own value of travel time. In the tricriterion case, each path is associated with a *region* in $\mathbb{R}^2_+$; each user chooses the path associated with the region containing their values of $VOATT$ and $VOTR$. Figure 5.8 shows these regions for the paths in Table 5.7.

Note that the vertices in Figure 5.7 correspond to regions in Figure 5.8, and vice versa; thus, these two graphs can be considered dual to each other. For instance, the point $(2, 2)$ in Figure 5.8 is adjacent to the regions corresponding to all four paths, indicating that when $VOATT = VOTR = 2$, these paths have equal costs. The corresponding region in Figure 5.7 is the plane segment P1-P2-P4-P3. One can verify that the vector $[\tau, E[t], Var[t]] = [2, 4, 4]$ is normal to this plane segment, suggesting that the plane segment represents

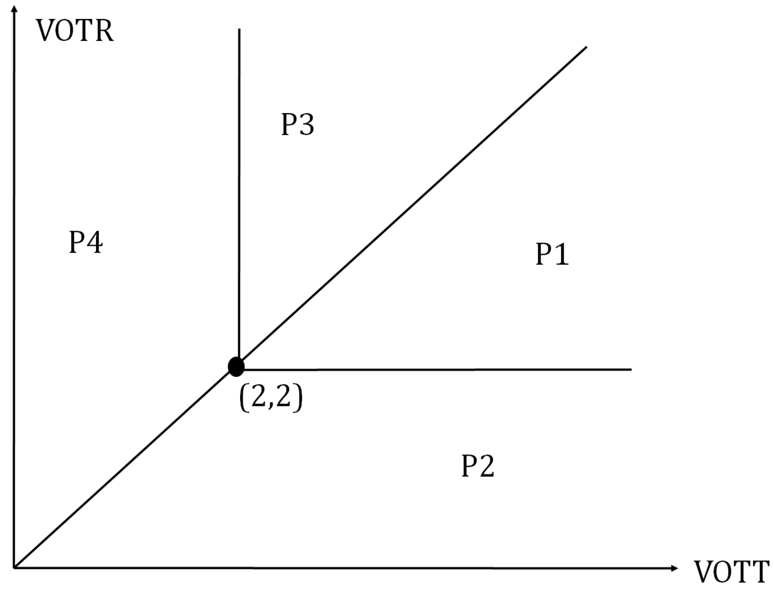


Figure 5.8: Regions of $VOTT$-$VOTR$ space corresponding to each efficient path (the "dual graph")

$(VOTT, VOTR) = (4/2, 4/2) = (2, 2)$, thus demonstrating the correspondence with the dual graph.

A general procedure for identifying efficient paths can now be described. Given a triangular plane segment with vertices $\{\mathbf{P^1}, \mathbf{P^2}, \mathbf{P^3}\}$, the corresponding values of $VOATT$ and $VOTR$ can be obtained from a vector normal to the segment (a normal vector can easily be found by taking the cross product of $\mathbf{P^2} - \mathbf{P^1}$ and $\mathbf{P^3} - \mathbf{P^1}$). The least-cost path π for these particular values of $VOATT$ and $VOTR$ are identified, along with the corresponding criterion vector $\mathbf{P}^\pi$; if this path is not already part of the efficient set, add it, and divide the plane segment into three new segments: $\{\mathbf{P^1}, \mathbf{P^2}, \mathbf{P}^\pi\}$, $\{\mathbf{P^1}, \mathbf{P^3}, \mathbf{P}^\pi\}$, and $\{\mathbf{P^2}, \mathbf{P^3}, \mathbf{P}^\pi\}$. These new plane segments are then recursively examined in the same way, until all efficient paths have been identified. It is worthwhile to update the dual graph at the same time, by eliminating the point corresponding to the original plane segment, adding the three points corresponding to the new plane segments, and updating the edges so that adjacent plane segments in the primal are directly connected in the dual.

One must be careful in initializing this algorithm. For the bicriterion case, it suffices to identify the least-cost and least-time paths, and use the slope of the line connecting their bicriterion vectors to begin the recursion. However, for the tricriterion case, this may cause difficulties due to tiebreaking. Continuing with the example in Table 5.7, if one initializes the algorithm with paths P1, P2, and P4 (least-mean time, least variance, and least-cost paths, respectively), path P3 (which is also a least variance path) will never be identified. This difficulty is avoided by creating three artificial tricriterion vectors $(M, 0, 0)$, $(0, M, 0)$, and $(0, 0, M)$, for a large scalar M, and using their convex hull as the initial plane segment.

Once the efficient paths have been generated, travelers must be assigned to each path based on their values of travel time and reliability. The number of travelers choosing a path is equal to the double integral of the joint density functions for $VOATT$ and $VOTR$, taken over the corresponding region in the dual graph. If the density function is difficult to integrate, an alternate method is to use a Monte Carlo method to generate points $(VOTT, VOTR)$, and applying a point-in-polygon algorithm (see, for instance, Preparata and Shamos, 1985, pp. 41–67) to identify the appropriate path.

Given (fixed) tolls, mean travel times, and travel time variances, this procedure can be used to assign trips to users with varying values of time and reliability. The equilibrium algorithms presented in Dial (1996, 1997) can then apply this assignment procedure repeatedly to find a user equilibrium with heterogeneous users (the remaining modifications to account for three criteria, rather than two, are straightforward). Following Dial (1999a), system-optimal arc flows and first-best tolls can then be found using this equilibrium procedure, by defining arc tolls as a function of arc flows:

$$\tau_{ij}(x_{ij}) = x_{ij}\left(\overline{VOATT}_{ij}\frac{dE[t_{ij}]}{dx_{ij}} + \overline{VOTR}_{ij}\frac{dVar[t_{ij}]}{dx_{ij}}\right) \tag{5.20}$$

where $\overline{VOATT}_{ij}$ and $\overline{VOTR}_{ij}$ are the mean values of average travel time and reliability for all users on arc (i,j). With this toll function, a tricriterion user equilibrium coincides with the system-optimum.

By accounting for variation in user preferences, the model more accurately represents traveler decision-making, and thus allows more accurate selection of toll levels.

5.3.5 Demonstration

The impacts of different information scenarios and users' valuation of reliability were studied using the Sioux Falls test network. The basic model developed in Sections 5.3.2 and 5.3.3 is applied — in the interest of space and clarifying the main impacts of uncertain supply, the extensions in the previous section are not considered here. The capacity on freeway arcs was made random, equal to its nominal value with probability 0.90, and reduced to a third of its nominal value with probability 0.10 (representing a major incident occurring one day out of ten). A \$10/hr value of travel time was assumed, for both $VOTT$ and $VOATT$.

In the absence of data to calibrate the $VOTR$ parameter directly, a rough estimate is made based on the results of Small et al. (2005), whose revealed preference data showed travelers were willing to spend \$19.56/hr to reduce the difference between 80th- and 50th-percentile travel times, compared to a \$21.46/hr value of travel time; with our $VOTT$ assumption, one expects that our travelers would pay \$9.11 for the same, the proportionate amount. Since their experiment most closely resembled the NI/UT scenario (although tolls changed dynamically, this was done in response to recurrent congestion, rather than incidents), the basic model was initially run with $VOTR = 0$ and the average travel time found. Assuming a normal distribution on trip travel time, the difference between the 80th- and 50th-percentile travel times for the NI/UT scenarios is 1.46 minutes, indicating that our travelers would pay \$0.22 to eliminate this uncertainty; with a variance of 2.97 minutes squared, this implies a \$0.074/min^2 $VOTR$.

Each of the solution methods in the previous section was applied to the appropriate information scenario. For comparison with heuristics H1 and H2 for the AR/UT scenario, the simulated annealing (SA) metaheuristic, developed by Kirkpatrick et al. (1983), was also applied to generate an approximately optimal toll vector; for SA, solution neighbors were obtained

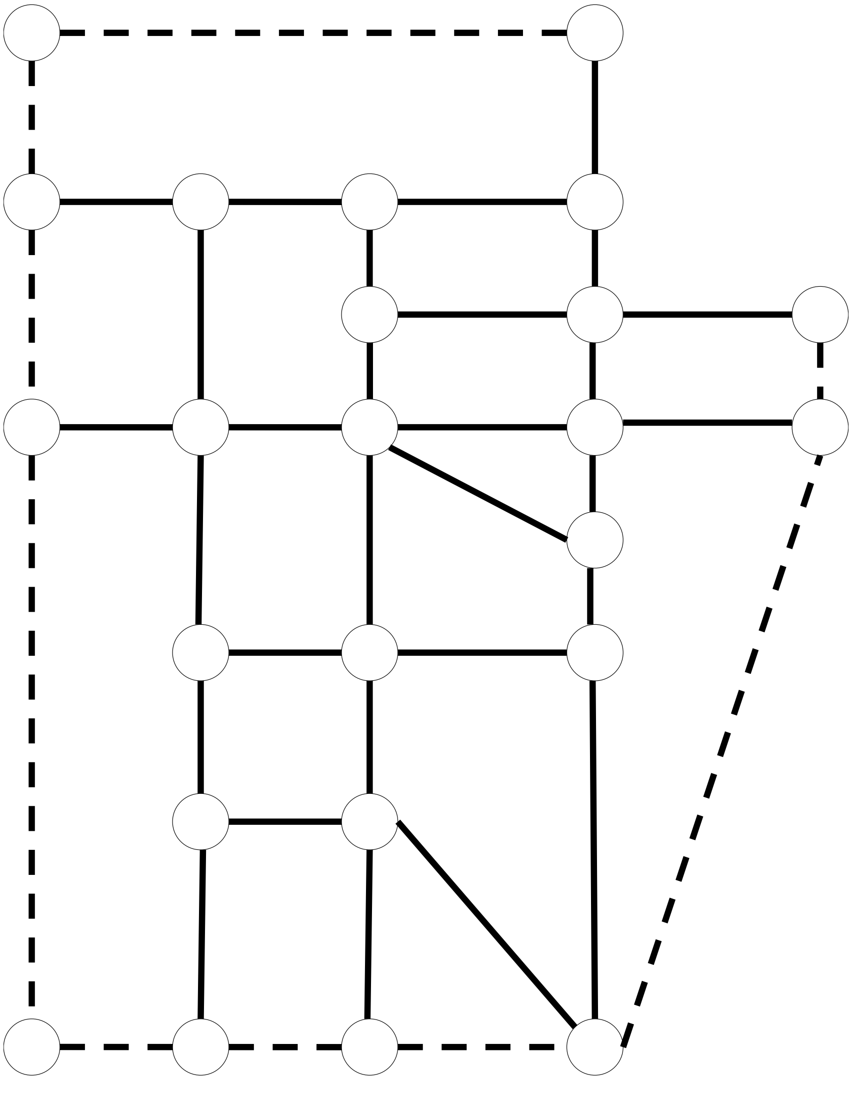


Figure 5.9: Sioux Falls test network; dashed lines indicate degradable (freeway) arcs.

Table 5.8: Comparison of average trip characteristics for all information scenarios

Users	Operator	Travel time (min)		Toll paid ($)		Run
		Mean	Std. Dev.	Mean	Std. Dev.	time (s)
NI	No Toll	23.60	0.74	0	0	—
	UT	23.55	0.70	9.82	0	5
AR	No Toll	21.85	0.94	0	0	—
	RT	20.33	0.53	5.74	0.03	11
	UT - H1	20.25	0.55	5.89	0.01	215
	UT - H2	20.24	0.55	5.90	0.01	552
	UT - SA	20.21	0.56	5.89	0.01	417

by perturbing arc tolls by up to fifty cents each. Additionally, for comparison, “no-toll’ scenarios were evaluated for the NI and AR user information scenarios.

Tables 5.8 and 5.9 compare the mean and standard deviation of trip durations and toll charges under the different information scenarios, along with the computation time needed for each solution method. Table 5.8 compares “average” trip characteristics; that is, the mean and variance of each traveler’s day-to-day travel times and tolls were first calculated, and averages of these values were taken across all travelers. Table 5.9 also lists the numerical value of the manager’s objective function, representing the total burden due to travel time and travel variability. Note that the standard deviations shown represent the variation seen over a period of many days — although travelers experience no uncertainty within a given day for the fully informed cases, there still is variation between days in their experienced travel times and toll expenses.

Several observations are apparent. First, providing users with pre-trip

Table 5.9: Comparison of system states for all information scenarios

Users	Operator	Total travel time (veh-hr)		Toll revenue ($ $\times 10^3$)	
		Mean	Std. Dev.	Mean	Std. Dev.
NI	No Toll	141836	4447	0	0
	UT	141536	4207	3542	0
AR	No Toll	131318	5652	0	0
	RT	121183	3184	2070	11
	UT - H1	122702	3306	2124	4
	UT - H2	122641	3305	2127	4
	UT - SA	122466	3313	2125	4

information on system conditions provides a substantial reduction in average travel times, on the order of ten to fifteen percent. Furthermore, with our assumptions on $VOTT$, $VOTR$, and system reliability, this benefit exceeds the benefit of congestion pricing. Second, marginal-cost tolls are higher when users do not have access to information on the network state. This occurs because the effect of a potential incident must always be incorporated into the toll price — since an incident result in large delays, a large toll is needed to correct the situation. Responsive tolling and providing users information allow more finesse: if users are aware of an incident, many will choose alternate routes on their own, even without a high toll; and responsive tolling allows levying a high toll only when warranted by an incident. A more mathematical reason is that the Pigouvian toll must include a term representing the marginal loss in reliability in addition to the marginal increase in average travel time, unless information is provided.

Third, congestion pricing has a greater impact in improving average travel times (as compared to the no-toll case) when users have information, but the marginal-cost tolls are much higher on average. For the responsive tolling scenarios, this occurs because the tolls on degraded arcs can be selectively increased, providing additional disincentive for using such arcs — without tolling, the increased travel times also discourage use of these arcs, but prices allow for an even greater reduction in total system travel time. Even for the unresponsive tolling scenarios, high tolls appear to be needed, perhaps to prevent users from "overcorrecting" when they learn of reduced capacity on their original path choice, creating additional congestion on a secondary route even as their own travel time decreases.

Finally, as is common with marginal-cost tolling, the levied tolls are greater in magnitude than the reduction in travel disutility. Nevertheless, since toll revenues are assumed to be returned to the public in some fashion, as long as the cost of implementing and administering the toll system is smaller than the reductions in disutility indicated in Table 5.9, a net social benefit still obtains.

For finding tolls in the AR/UT case, simple averaging (H1) appears to work just as well as the alternate heuristic H2, or simulated annealing, while requiring somewhat less computation time; it would be interesting to see whether this result occurs in larger networks as well.

More insight is obtained by examining differences among individual arc conditions under these different scenarios. As base cases for comparison, Figures 5.10 and 5.11 respectively show average arc volume-to-capacity (v/c) ratios for the no-information and adaptive-routing cases when no tolls are present. Note that although arterials are substantially congested (almost all have $v/c > 1$), freeway arcs are generally underutilized, as a result of their uncertain capacity. When travelers can choose their routes flexibly, freeways are used slightly more, but not significantly so. This is a result of the local nature of the information provided in this problem: users are not willing to detour to use a freeway, as they might do if they could learn its state farther in advance.

Figures 5.12, 5.13, and 5.14 show the change in v/c ratios when tolls are applied in the NI/UT, AR/RT, and AR/UT cases, compared with the untolled base cases. For the uninformed (NI/UT) scenario, the general effect of the tolls is to shift flow onto less congested routes, such as the northern freeways. Increased freeway usage is somewhat apparent in the informed (AR/RT and AR/UT) scenarios, where the combination of information provision and tolls are effective in persuading travelers to use less-congested freeways to the north and west, leading to gains in system-wide operations.

5.3.6 Conclusion

This section considers first-best pricing problems in the presence of network uncertainty and user valuation of travel time reliability, allowing for adaptive route choice. As with any stochastic model, the question of information is key. Four information provision scenarios are developed, accounting for both network managers and users, although one of these is shown to be a special case of another: from the standpoint of encouraging system-optimal

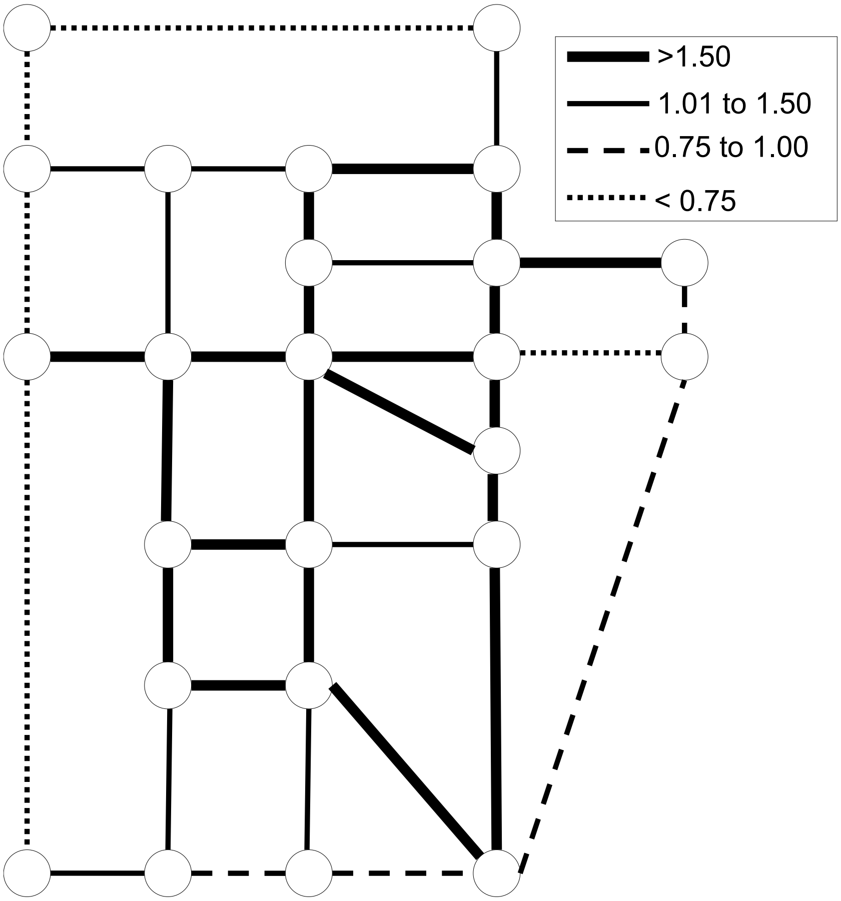


Figure 5.10: Untolled volume-to-capacity ratios for no-information case

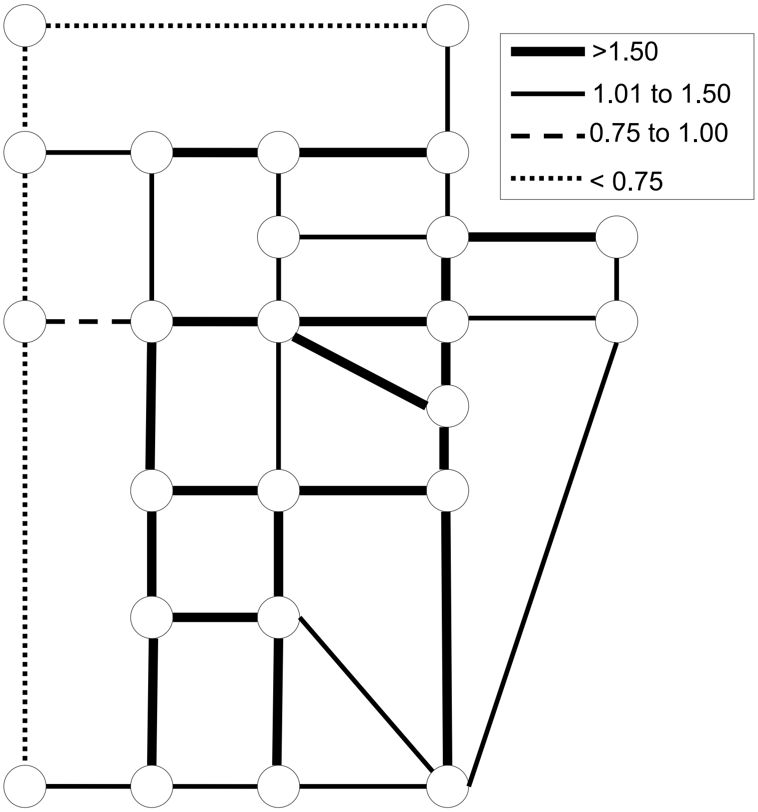


Figure 5.11: Untolled volume-to-capacity ratios for full-information case

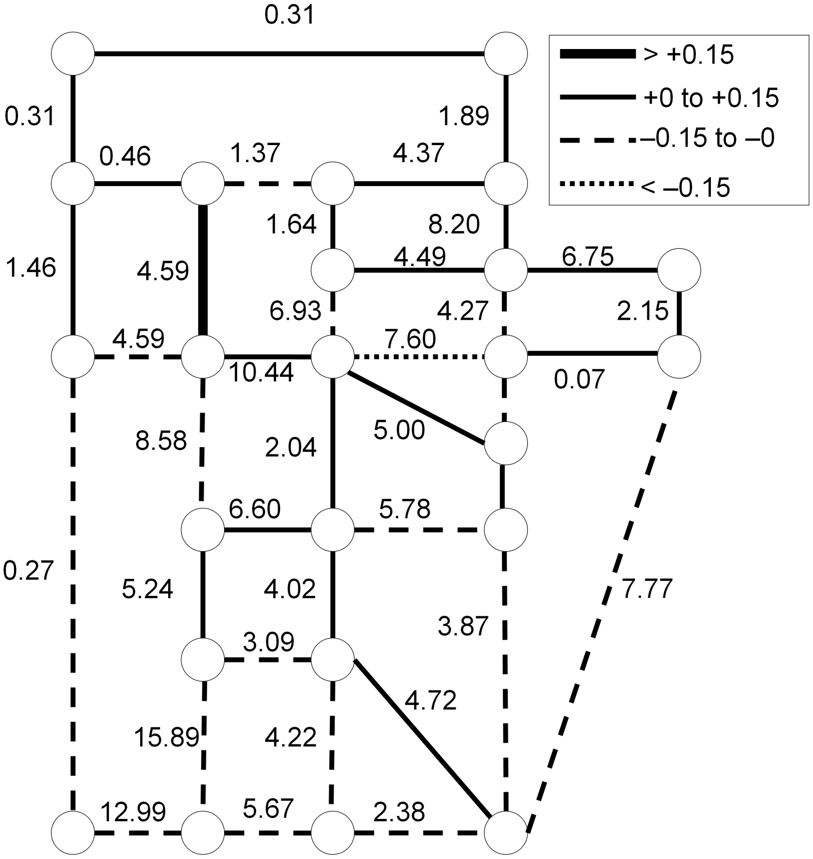


Figure 5.12: Change in v/c ratios under tolling (NI/UT); average arc tolls shown in dollars

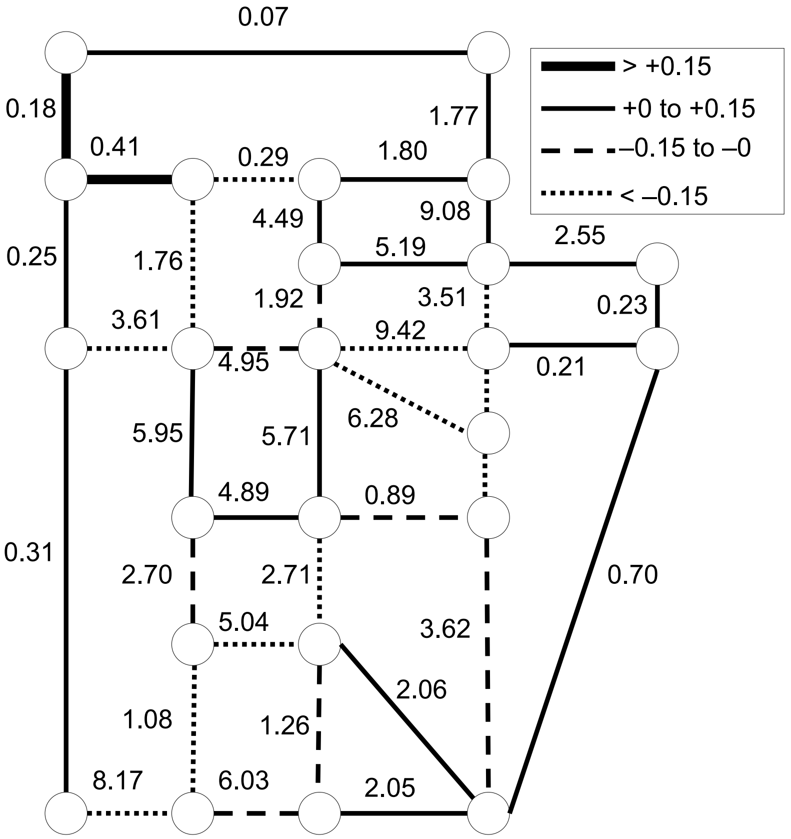


Figure 5.13: Change in v/c ratios under tolling (AR/RT); average arc tolls shown in dollars

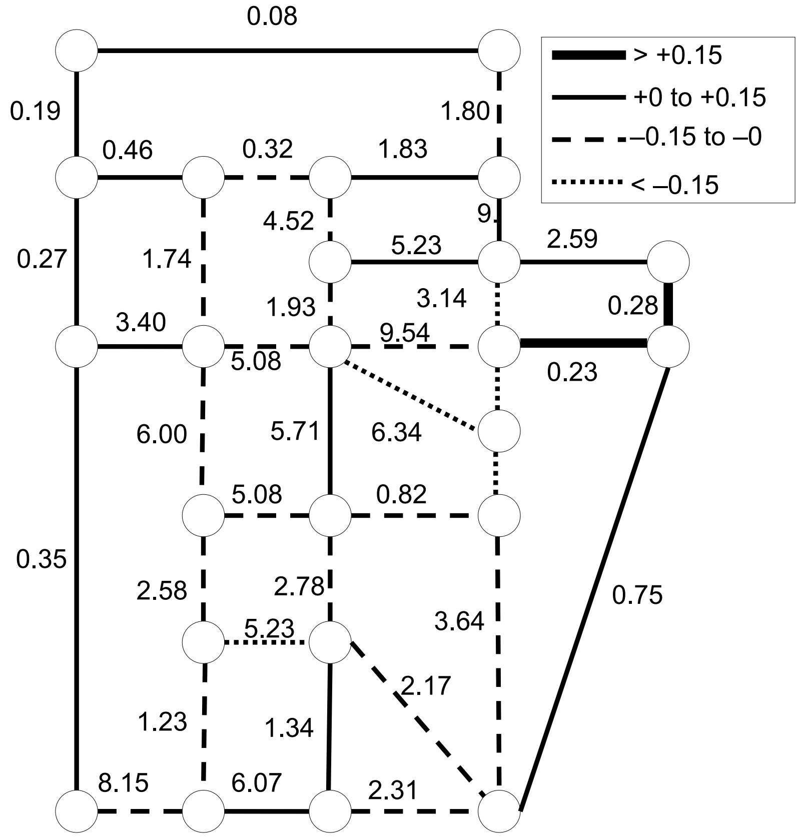


Figure 5.14: Change in v/c ratios under tolling (AR/UT); average arc tolls shown in dollars

behavior, there is no value in varying tolls if users do not learn of the network realization. Solution methods are presented for each of these scenarios, which were then tested on the Sioux Falls network.

Perhaps the most noteworthy conclusion is that unresponsive tolls must be set higher than responsive tolls, since the network manager must always hedge against rare events to ensure the convergence of system optimal and user optimal behavior. This suggests that unresponsive tolling should not be used to address nonrecurring congestion, but instead be limited to recurring, predictable congestion. Responsive tolls, however, do not suffer from this weakness, assuming full information on behalf of network managers and travelers.

5.4 Network Design

The equilibrium model in Chapter 3 already provides the means to evaluate a wide variety of policies aimed at improving system conditions, including lane addition, signal cycle adjustment, and construction of new arcs, yielding estimates of these policies' benefits. These estimates can then be used to prioritize or select potential improvements. However, at times, more specific guidance is desired: instead of simply evaluating previously-conceived improvements, the network design problem generates a set of improvements which optimizes an objective such as TSTT.

Network design models come in many varieties; in transportation networks, they often take the form of mathematical programs with equilibrium constraints (MPECs), where the improvements are treated as increases in the capacity parameter of the arc performance function. In this section, we assume that capacity can be continuously added to arcs; let z_{ij} denote the additional capacity on arc (i, j) for *all* of its associated states — that is, more capacity cannot be added to one state than another. Further assuming that the marginal cost of providing additional capacity is constant for all arcs and existing capacity levels, we define the budget B indicating the total amount of additional capacity to be added to arcs, so we have $\mathbf{z} \cdot \mathbf{1} \leq B$.

In practice, arc improvements are often discrete; adding half of a travel lane certainly does not provide half the additional capacity of a full lane. This is not restrictive. As MPECs are difficult to solve, researchers often apply a battery of metaheuristics such as simulated annealing, genetic algorithms, or tabu search; all of these employ some discretization of the feasible space, either implicitly or explicitly. In fact, a coarser discretization will greatly reduce the running time of these solution methods.

Thus, we define the network design problem NL-UER-NDP as follows:

$$\min_{\mathbf{x},\mathbf{y},\mathbf{z}} \quad TSTT = \sum_{(i,j)\in A} \sum_{s\in S_{ij}} x_{ij}^s t_{ij}^s(x_{ij}^s, z_{ij}) \tag{5.21}$$

$$\text{s.t.} \quad \sum_{(u,v)\in D} \sum_{q\in Q_{uv}} \sum_{\pi\in\Pi_{uv}} F_q(\mathbf{y},\mathbf{z})(y_q^\pi - \hat{y}_q^\pi) \leq 0 \qquad \forall \hat{\mathbf{y}} \in Y \tag{5.22}$$

$$x_{ij}^s = \sum_{(u,v)\in D} \sum_{q\in Q_{uv}} \sum_{\pi\in\Pi_{uv}} a_{\pi,[ijs]} y_q^\pi \sum \tag{5.23}$$

$$\sum_{(i,j)\in A} z_{ij} \leq B \tag{5.24}$$

$$x_{ij}^s \geq 0 \qquad \forall (i,j) \in A, s \in S_{ij} \tag{5.25}$$

$$\mathbf{y} \in Y \tag{5.26}$$

$$z_{ij} \geq 0 \qquad \forall (i,j) \in A \tag{5.27}$$

$$\tag{5.28}$$

where

$$Y = \left\{ \mathbf{y} \in \mathbb{R}^{\sum_{(u,v)\in D} |Q_{uv}||\Pi_{uv}|} : y_q^\pi \geq 0, \sum_{\pi\in\Pi_{uv}} y_q^\pi = d_{uv}^q \right\} \tag{5.29}$$

is the set of feasible policy flows, reflecting nonnegativity and demand satisfaction. The objective function (5.21) minimizes the total system travel time. Note that we are not minimizing the sum of traveler disutility, because of inherent difficulties in making different user classes' disutility functions commensurate, and because disutility is unobservable in practice. Constraint (5.22) expresses the equilibrium condition as a variational inequality, incorporating the capacity enhancements **z**; it is the presence of this constraint that makes solution of NL-UER-NDP difficult. The remaining constraints are straightforward: (5.23) maps policy flows to arc-state flows, (5.24) reflects the budget constraint, and the remainder address nonnegativity and demand feasibility.

As suggested previously, many researchers have studied similar network design problems, and a comprehensive literature review is beyond the scope of this section. However, several key works are worth mentioning. Historically, Abdulaal and LeBlanc (1979) and Marcotte (1983) were among the first to consider equilibrium-based transportation network design. The former applies the Hooke-Jeeves method to approximate the derivative of the objective function; the latter develops an algorithm based on a row generation approach involving transformation of the variational inequality constraint. Both of these methods are heuristic in nature. Yang and Bell (1998) and Karoonsoontawong (2006) provide fuller overviews of the development of transportation network design problems since these initial efforts. A small subset of the solution methods applied includes branch-and-bound, equilibrium decomposed optimization, Bard's algorithm, Powell's method, simulated annealing, genetic algorithms, and tabu search. However, the only instance of this type of network design model with adaptive routing is in Unnikrishnan (2008), approximately solved using two types of genetic algorithm. This section builds on this work by also considering nonlinear disutility functions.

The aim of this section is not add new solution methods to the plethora of existing methods for this type of problem, but instead to simply state the NL-UER version and consider its numerical properties on the Sioux Falls network. Capacity can be added in increments of 500 veh/hr, and the budget allows fifty such increments to be made across the network. A single user class is assumed, with a quadratic disutility function. Simulated annealing is used to find near-optimal arc improvements; as before, the approach of Chiang and Russell (1996) is used to determine the cooling schedule.

An initial solution is generated with the following heuristic approach. Assuming BPR-type delay functions and temporarily fixing the arc-state flows at their current levels, the marginal decrease obtained by increasing the capacity of arc (i, j) is

$$
\begin{aligned}
-\frac{\partial TSTT}{\partial c_{ij}} &= \sum_{s \in S_{ij}} x_{ij}^s \frac{\partial}{\partial c_{ij}} t_{ij}^s(x_{ij}^s) && (5.30)\\
&= \sum_{s \in S_{ij}} x_{ij}^s \left[t_0^s \alpha \left(\frac{x_{ij}^s}{p_{ij}^s c_{ij}^s} \right)^{\beta} (-\beta c_{ij}^s)^{-(\beta+1)} \right] && (5.31)\\
&= \sum_{s \in S_{ij}} p_{ij}^s \left(\frac{x_{ij}^s}{p_s c_{ij}^s} \right)^{\beta+1} && (5.32)
\end{aligned}
$$

These are used to determine the initial distribution of improvements: the first increment is assigned to the arc (i, j) maximizing (5.32), increasing its capacity and reducing the value of (5.32) for this arc. The second increment is again assigned to the arc for which the sum (5.32) is greatest, accounting for the change in capacity (but not flow) induced by the first increment; this procedure is repeated until all of the increments have been assigned.

The results of this procedure can be seen in Table 5.10. The left side of the table indicates the initial network conditions through the demand-to-capacity ratio x/c and arc delays for the no-incident and incident states. The right side indicates the conditions after implementing the capacity increases marked in the center column z. As a result of these improvements, TSTT is reduced by nearly forty percent from 4.8×10^7 vehicle-minutes to 3.0×10^7.

These gains are probably unrealistic because of the excessive x/c ratios seen on some of the arcs in the initial state. For instance, arc $(8, 6)$ initially has a demand-to-capacity ratio in excess of four. Due to the convex nature of BPR functions, additional demand in excess of capacity results in very sharp increases in delay, in excess of 250 minutes — clearly a physical impossibility, no matter how severe the queueing and congestion.

Thus, much of the congestion relief which appears when solving NDP may simply be restoring delays to realistic values; nevertheless, such locations are likely to be congested anyway, and improving bottlenecks here may have impacts throughout the network.

Table 5.10: Results from network design on Sioux Falls

Arc	$(x/c)^{NI}$	$(x/c)^{IP}$	t^{NI}	t^{IP}	z	$(x/c)^{NI}$	$(x/c)^{IP}$	t^{NI}	t^{IP}
(1,2)	0.24	0.01	6	18	0	0.23	0.01	0	18
(1,3)	0.28	0.16	4	12	0	0.28	0.16	0	12
(2,1)	0.03	0.00	6	18	0	0.03	0	0.00	18
(2,6)	1.42	1.00	5	17	0	1.42	1.00	0	15
(3,1)	0.47	0.36	4	12	0	0.28	0.19	0	12
(3,4)	0.66	0.21	4	12	0	0.63	0.19	0	12
(3,12)	0.37	0.28	4	12	0	0.37	0.28	0	12
(4,3)	0.78	0.06	4	12	0	0.41	0.02	0	12
(4,5)	1.15	0.78	2	6	0	0.81	0.51	0.00	6
(4,11)	1.66	0.00	6	18	0	1.65	0.00	0	18
(5,4)	0.62	0.54	2	6	0	0.57	0.50	6	6
(5,6)	3.16	0.24	4	12	500	2.21	0.15	4	12
(5,9)	0.54	0.00	5	15	0	0.53	0.00	6	15
(6,2)	0.31	0.00	5	15	0	0.46	0.04	5	15
(6,5)	0.76	0.02	4	12	0	1.19	0.07	4	12
(6,8)	3.69	3.32	2	116	1000	3.1	2.78	4	6
(7,8)	2.16	0.20	3	9	0	2.41	0.25	4	9
(7,18)	0.92	0.60	2	6	0	1.03	0.70	4	6
(8,6)	4.49	4.06	2	250	2000	2.43	2.09	2	6
(8,7)	1.41	0.04	3	9	0	1.84	0.07	6	9
(8,9)	0.15	0.00	10	30	0	0.15	0.00	2	30
(8,16)	1.96	0.00	5	15	0	2.55	0.02	4	15
(9,5)	0.77	0.06	5	15	0	0.72	0.05	5	15
(9,8)	0.12	0.00	10	30	0	0.14	0.00	5	30
(9,10)	1.18	0.72	3	9	0	1.14	0.72	4	9
(10,9)	0.85	0.48	3	9	0	0.74	0.44	2	9
(10,11)	1.5	0.22	5	15	0	0.76	0.01	3	15
(10,15)	1.29	0.00	6	18	0	1.28	0.00	2	18
(10,16)	5.18	0.04	4	12	4000	2.88	0.07	2	12
(10,17)	0.34	0.00	8	24	0	0.34	0.00	3	24
(11,4)	3.01	0.07	6	18	1000	3.87	0.24	10	18
(11,10)	0.98	0.04	5	15	0	1.36	0.17	5	15
(11,12)	1.49	0.28	6	18	0	1.32	0.05	5	18

Continued on next page...

Table 5.10 – Continued

Arc	$(x/c)^{NI}$	$(x/c)^{IP}$	t^{NI}	t^{IP}	z	$(x/c)^{NI}$	$(x/c)^{IP}$	t^{NI}	t^{IP}
(11,14)	4.41	1.54	4	22	2500	2.9	0.99	10	12
(12,3)	0.62	0.04	4	12	0	0.53	0.03	3	12
(12,11)	1.52	0.00	6	18	0	1.49	0.01	3	18
(12,13)	0.61	0.44	3	9	0	0.58	0.42	5	9
(13,12)	0.5	0.34	3	9	0	0.3	0.15	6	9
(13,24)	3.04	0.30	4	12	500	2.76	0.28	4	12
(14,11)	4.17	2.00	4	41	2500	2.89	1.43	8	12
(14,15)	0.83	0.01	5	15	0	2.43	0.05	6	15
(14,23)	1.4	0.63	4	12	0	2.61	1.00	5	12
(15,10)	0.79	0.02	6	18	0	0.82	0.15	6	18
(15,14)	2.24	0.17	5	15	0	1.78	0.04	4	15
(15,19)	1.14	0.56	3	9	0	1.46	0.7	4	9
(15,22)	2.98	1.13	3	11	0	2.54	0.93	6	9
(16,8)	1.59	0.01	5	15	0	3.57	0.05	3	15
(16,10)	3.33	0.03	4	12	1000	3.74	0.16	3	12
(16,17)	3.4	1.15	2	8	500	4.44	2.05	4	6
(16,18)	0.15	0.00	3	9	0	1.05	0.05	4	9
(17,10)	0.22	0.01	8	24	0	0.19	0.01	5	24
(17,16)	4.96	3.11	2	90	2500	2.69	1.62	4	6
(17,19)	3.24	2.98	2	77	0	1.12	0.89	6	6
(18,7)	1.05	0.31	2	6	0	1.42	0.58	5	6
(18,16)	0.88	0.06	3	9	0	0.91	0.06	3	9
(18,20)	0.64	0.01	4	12	0	0.88	0.02	3	12
(19,15)	1.1	0.10	3	9	0	0.58	0.05	5	9
(19,17)	3.54	2.63	2	49	500	3.12	2.29	4	6
(19,20)	0.55	0.03	4	12	0	1.07	0.08	2	12
(20,18)	0.64	0.10	4	12	0	0.81	0.16	3	12
(20,19)	0.9	0.03	4	12	0	1.04	0.06	8	12
(20,21)	1.94	0.00	6	18	0	1.93	0.01	2	18
(20,22)	1.55	0.03	5	15	0	1.52	0.04	2	15
(21,20)	0.74	0.01	6	18	0	1.94	0.01	2	18
(21,22)	2.13	1.10	2	7	0	3.87	2.20	3	6
(21,24)	5.14	1.15	3	11	3500	3.07	0.33	4	9
(22,15)	2.54	0.14	3	9	0	2.80	0.40	3	9

Continued on next page...

Table 5.10 – Continued

Arc	$(x/c)^{NI}$	$(x/c)^{IP}$	t^{NI}	t^{IP}	z	$(x/c)^{NI}$	$(x/c)^{IP}$	t^{NI}	t^{IP}
(22,20)	1.47	0.03	5	15	0	1.49	0.05	2	15
(22,21)	3.72	0.97	2	7	1000	2.99	0.80	4	6
(22,23)	2.04	0.02	4	12	0	2.02	0.03	4	12
(23,14)	2.48	0.14	4	12	0	2.76	0.23	4	12
(23,22)	2.81	0.02	4	12	0	3.21	0.07	6	12
(23,24)	2.08	1.76	2	15	0	2.73	2.33	5	6
(24,13)	1.63	0.02	4	12	0	3.87	0.37	6	12
(24,21)	4.09	0.21	3	9	2000	2.90	0.18	2	9
(24,23)	1.34	0.71	2	6	0	1.95	0.90	3	6

Chapter 6

Conclusion

6.1 Summary and Implications

This dissertation developed a suite of models and algorithms which can be used to trace the impacts of operational, supply-side uncertainty from their causes, to their effects, and to their impact on agency policies for improving network conditions. As discussed in Chapter 1, nonrecurring congestion constitutes a very significant portion of travel delay, and simply replacing uncertain parameters with their expected values leads to systematic error both in describing the network state, and in recommending improvements that can be made. Thus, there is a need to develop methods for incorporating operational uncertainty into transportation planning models, and it is to this need that this dissertation speaks. Two byproducts of uncertainty — nonlinear risk attitudes and re-routing due to information provision — are made an integral part of this dissertation's contributions. The former has repeatedly shown to be important in decision making in transportation systems, while inclusion of the latter allows evaluation of a broad spectrum of ITS technologies at the network level, in a way that has heretofore been impossible.

Chapter 2 focused on the causes of uncertainty, and provided a statistical procedure for estimating probability distributions of operational metrics (such as travel speed or capacity) in locations where no data is available. This procedure develops a regression model, using locations where data *is* available to relate roadway geometry, physical location, and other quantities to the operational metrics in question. Demonstrations were provided for estimating speed distributions, for estimating roadway capacity, and for estimating capacity degradation under incidents. Analytically, closed-form expressions were provided for expected delay to travelers arriving at traffic signals under different indications, and for incorporating uncertainty in travel demand into these estimations.

Chapter 3 emphasized the effects of uncertainty on individual and collective traveler behavior, and contained the major methodological contributions of the dissertation. Individual travel behavior was represented using a modified online shortest path algorithm which accounts for both *en route* information and nonlinear user preferences. Under certain conditions, *contretemps* can arise, where the path a traveler follows can contain arbitrarily many cycles even under an optimal routing policy. Even though this behavior is rare in realistic networks, its possibility requires certain algorithmic precautions to be taken to ensure finite termination. The dissertation takes the approach of forcing trips to end before a given time horizon, and provides an analytical bound on the error introduced by doing so. The good news is that, for almost any disutility function that has been considered in the literature, this bound becomes arbitrarily small as the time horizon becomes longer.

Collective behavior was modeled using an equilibrium approach, where the equilibrium exists among routing policies, rather than simple paths, due to the possibility of online information. Through a suitable transformation to an asymmetric deterministic user equilibrium problem, key theoretical propeties were established, including existence of this equilibrium under weak conditions, and uniqueness under the slightly stronger condition that disutility functions are monotone. This equilibrium can be expressed as the solution of a variational inequality, a class of problems with a well-developed theory and multiple solution methods. An algorithm converging to such an equlibrium was also provided. Unique among online equilibrium algorithms, a more efficient policy loading procedure was developed, allowing all policies with a common destination to be loaded on the network simultaneously.

These algorithms function in time-expanded networks. Although significantly larger than the original networks on which they are built, their acyclic nature allows certain efficiencies in network algorithms which keep the computational burden from growing too large. Use of time-expanded networks carries other side benefits, such as the trivial incorporation of departure time choice by including a self-arc from each origin to itself.

Chapter 5 gave three example applications of how this modeling approach can be fruitfully applied. The first, locating information provision, is impossible to accomplish without adaptive routing and equilibrium models. Solution methods were developed for three problems of this type: identifying the best locations to provide information to an individual driver (allowing preparation of adaptive driving directions in a format easily understood by drivers); identifying the best locations for multiple drivers, but where congestion can be ignored (as in regional freight models in rural areas which are subject to inclement weather); and identifying the best locations for multiple drivers where congestion cannot be ignored (as when locating VMSs in urban areas). For the first two problems, a network contraction procedure allows feasible solutions to be evaluated extremely rapidly.

The second application, congestion pricing, has existed for decades; however, the technology for adaptive pricing is more recent. The online equilibrium model developed here allows first-best adaptive prices to be calculated; this could not be accomplished using past equilibrium models. This is compared with application of static prices, and in scenarios where travelers do not have access to information. Finally, the canonical network design problem is considered, showing how the equilibrium model can provide additional insight on traditional alternatives analysis, making recommendations which can account for travelers' risk preferences and travel information.

In addition to the contributions made in the individual chapters, a major contribution of this dissertation is the combination of these models with a common set of assumptions and within a single unifying theoretical framework.

6.2 Future Work

Many opportunities exist to extend the models presented here in both the theoretical and practical directions. One natural extension is considering combinations of the three mitigation strategies discussed in Chapter 5 to see if multiple strategies can act synergistically; for instance, adding capacity to links incident to information nodes may yield benefits superior to capacity addition or information provision alone. Other mitigation strategies involving ITS can also be included in such an analysis.

Determining the appropriate form that travelers' disutility functions take remains a major challenge which deserves much attention. Although economists and demand modelers have begun to explore the question of how best to represent traveler risk preferences, there is still no consensus on how route choice under uncertainty should best be represented. For practical applications, this question is key, especially detecting origins or destinations where strong heterogeneity in disutility functions may be observed, such as airports.

Relaxing several of the modeling assumptions can enhance the realism of these models. Some of these, such as the independence assumption, seem necessary to ensure reasonable running time; for instance, it has been proved that relaxing the "reset" assumption of independence between successive visits to an arc results in NP-hard online routing problems. Still, the development of practical heuristics for these scenarios would be worthwhile. Likewise, allowing state probabilities to depend on flow can allow better modeling of phenomena such as incidents.

From the standpoint of realism, perhaps the biggest shortcoming of the models presented here are their quasi-static nature, in which congestion does not vary with time. The time-expanded networks used in the individual and collective routing models provide a natural starting point for such a model, but realistic representation of dynamic traffic requires a move away from link performance functions to a more refined traffic flow model. Accomplishing this is certainly nontrivial, especially in a manner which does not require extreme amounts of computation time. Another approach is to integrate the routing model with a mesoscopic traffic simulator, which carries a related set of difficulties. In the latter case, a fundamental change must necessarily be made to the spatiotemporal independence assumption adopted throughout this dissertation.

Still, although these models can be improved in multiple directions, they constitute significant advances in representing uncertainty, traveler behavior, and network improvement in a consistent and interdependent manner. With continued research effort and practical experience, these models can help speak to the challenges transportation professionals face, through greater realism, and by allowing the effects of innovative technologies to be studied in a quantitative, rigorous manner.

References

Abbas, M. and P. McCoy (1999). Optimizing variable message sign locations on freeways using genetic algorithms. Presented at the 78th Annual Meeting of the Transportation Research Board, Washington, DC.

Abdulaal, M. and L. J. LeBlanc (1979). Continuous equilibrium network design models. *Transportation Research Part B 13B*, 19–32.

Ahuja, R. K., T. Magnanti, and J. Orlin (1993). *Network Flows*. Englewood, NJ: Prentice-Hall.

Akamatsu, T. and M. Kuwahara (1988). Optimal road pricing under stochastic user equilibrium. *Proceedings of the Japan Society of Civil Engineers 389/IV-8*, 121–129.

Altman, E. and L. Wynter (2004). Equilibrium, games, and pricing in transportation and telecommunication networks. *Networks and Spatial Economics 4*, 7–21.

Aron, M., M. Ellenberg, P. Fabre, and P. Veyre (1994). Weather related traffic management. In *Towards an Intelligent Transport System: Proceedings of the First World Congress on Applications of Transport Telematics and Intelligent Vehicle-Highway Systems*, Volume 3, Brussels, Belgium, pp. 1089–1096.

Asakura, Y. and M. Kashiwadani (1991). Road network reliability caused by daily fluctuations of traffic flow. In *Proceedings of the 19th PTRC Summer Annual meeting*, Brighton, pp. 73–84.

Avineri, E. and J. N. Prashker (2003). Sensitivity to uncertainty: Need for a paradigm shift. *Transportation Research Record 1854*, 90–98.

Bar-Gera, H. (2009). Transportation test problems. Website: `http://www.bgu.ac.il/~bargera/tntp/`. Accessed April 28, 2009.

Bates, J., J. Polak, P. Jones, and A. Cook (2001). The valuation of reliability for personal travel. *Transportation Research Part E 37*(2–3), 191–229.

Beckmann, M. J., C. B. McGuire, and C. B. Winston (1956). *Studies in the Economics of Transportation*. Connecticut: Yale University Press.

Bell, M. G. H., C. Cassir, Y. Iida, and W. H. K. Lam (1999). A sensitivity based approach to network reliability assessment. In *Proceedings of the 14th International Symposium on Transportation and Traffic Theory*, Jerusalem, pp. 283–300.

Ben-Akiva, M., A. de Palma, and I. Kaysi (1991). Dynamic network models and driver information systems. *Transportation Research Part A 25*(5), 251–266.

Box, G. and D. Cox (1964). An analysis of transformations. *Journal of Royal Statistical Society B 26*, 211–246.

Boyles, S. D. (2006). Reliable routing with recourse in stochastic, time-dependent transportation networks. Master's thesis, The University of Texas at Austin.

Boyles, S. D. and S. T. Waller (2007a). The impact of nonlinear objective functions in online shortest paths. Presented at the 86th Annual Meeting of the Transportation Research Board, Washington, DC.

Boyles, S. D. and S. T. Waller (2007b). Online routing with nonlinear disutility functions with arc cost dependencies. Presented at the Sixth Triennial Symposium on Transportation Analysis (TRISTAN VI), Phuket, Thailand.

Boyles, S. D. and S. T. Waller (2007c). A stochastic delay prediction model for real-time incident management. *ITE Journal 77*(11), 18–24.

Braess, D. (1969). Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung 12*, 258–268.

Brownstone, D. and K. A. Small (2005). Valuing time and reliability: assessing the evidence from road pricing demonstrations. *Transportation Research Part A 39*(4), 279–293.

Chabini, I. (1999). Discrete dynamic shortest path problems in transportation applications: complexity and algorithms with optimal run time. *Transportation Research Record 1645*, 170–175.

Chan, K. S. and W. H. K. Lam (2005). Impact of road pricing on the network reliability. *Journal of the Eastern Asia Society for Transportation Studies 6*, 2060–2075.

Chapra, S. C. and R. P. Canale (2002). *Numerical Methods for Engineers* (4th ed.). New York, NY: McGraw Hill.

Chen, A., M. Tatineni, D. H. Lee, and H. Yang (2000). Effect of route choice models on estimating network capacity reliability. *Transportation Research Record 1733*, 63–70.

Chiang, W.-C. and R. A. Russell (1996). Simulated annealing metaheuristisc for the vehicle routing problem with time windows. *Annals of Operations Research 63*, 3–27.

Chiu, Y.-C. and N. Huynh (2007). Location configuration design for dynamic message signs under stochastic incident scenarios. *Transportation Research Part C 15*(1), 333–50.

Chiu, Y.-C., N. Huynh, and H. Mahmassani (2001). Determining optimal locations for VMS's under stochastic incident scenarios. Presented at the 80th Annual Meeting of the Transportation Research Board, Washington, DC.

Clark, S. and D. Watling (2005). Modelling network travel time reliability under stochastic demand. *Transportation Research Part B 39*(2), 119–140.

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

de Lapparent, M., A. de Palma, and C. Fontan (2002). Nonlinearities in the valuation of travel attributes. In *Proceedings of the Professional Transportation Research Part Conference*, Cambridge, United Kingdom.

de Palma, A. and R. Lindsey (1998). Information and usage of congestible facilities under different pricing regimes. *Canadian Journal of Economics 31*(3), 666–692.

de Palma, A. and N. Picard (2005). Route choice decision under travel time uncertainty. *Transportation Research Part A 39*, 295–324.

Dial, R. B. (1996). Bicriterion traffic assignment: basic theory and elementary algorithms. *Transportation Science 30*, 93–111.

Dial, R. B. (1997). Bicriterion traffic assignment: efficient algorithms plus examples. *Transportation Research Part B 31*(5), 357–379.

Dial, R. B. (1999a). Network-optimized road pricing: Part I: a parable and a model. *Operations Research 47*(1), 54–64.

Dial, R. B. (1999b). Network-optimized road pricing: Part II: algorithms and examples. *Operations Research 47*(2), 327–336.

Du, Z. and A. Nicholson (1997). Degradable transportation systems: Sensitivity and reliability analysis. *Transportation Research Part B 31*(3), 225–237.

Duthie, J. C. (2005). Robust transportation network analysis with uncertain and correlated long-term origin-destination demands. Master's thesis, The University of Texas at Austin.

Eiger, A., P. B. Mirchandani, and H. Soroush (1985). Path preferences and optimal paths in probabilistic networks. *Transportation Science 19*(1), 75–84.

Emmerink, R. H. M., E. T. Verhoef, P. Nijkamp, and P. Rietveld (1996). Endogenising demand for information in road transport. *Annals of Regional Science 30*(2), 201–222.

Eswaran, P. K., A. Ravindran, and H. Moskowitz (1989). Algorithms for nonlinear integer bicriterion problems. *Journal of Optimization Theory and Applications 63*(2), 261–279.

Fan, Y., R. Kalaba, and J. Moore (2005). Arriving on time. *Journal of Optimization Theory and Applications 127*(3), 497–513.

Fan, Y. and Y. Nie (2006). Optimal routing for maximizing the travel time reliability. *Networks and Spatial Economics 6*, 333–344.

Floyd, R. W. (1962). Algorithm 97: Shortest path. *Communications of the ACM 5*(6), 345.

Frank, M. and P. Wolfe (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly 3*, 95–110.

Fu, L. and L. R. Rilett (1998). Expected shortest paths in dynamic and stochastic traffic networks. *Transportation Research Part B 32*(7), 499–516.

Fudenberg, D. and D. K. Levine (1998). *The Theory of Learning in Games.* Cambridge, MA: MIT Press.

Gabriel, S. A. and D. Bernstein (2000). Nonadditive shortest paths: subproblems in multi-agent competitive network models. *Computational and Mathematical Organization Theory 6*(1), 29–45.

Gao, S. (2005). *Optimal Adaptive Routing and Traffic Assignment in Stochastic Time-Dependent Networks.* Ph. D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Gao, S. and I. Chabini (2006). Optimal routing policy problems in stochastic time-dependent networks. *Transportation Research Part B 40*(2), 93–122.

Garib, A., A. E. Radwan, and H. Al-Deek (1997). Estimating magnitude and duration of incident delays. *Journal of Transportation Engineering 123*, 456–486.

Geoffrion, A. M. (1968). Proper efficiency and the theory of vector maximization. *Journal of Mathematical Analysis and Applications 22*, 618–630.

Golob, T., W. Recker, and J. Leonard (1987). An analysis of the severity and incident duration of truck-involved freeway accidents. *Accident Analysis and Prevention 19*, 375–395.

Golob, T. F. and W. W. Recker (2003). Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. *Journal of Transportation Engineering 129*(4), 342–353.

Haimes, Y. Y., L. S. Lasdon, and D. A. Wismer (1971). On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE Transactions on Systems, Man, and Cybernetics 1*(3), 296–297.

Hall, R. W. (1986). The fastest path through a network with random time-dependent travel times. *Transportation Science 20*(3), 182–188.

Hamdouch, Y., P. Marcotte, and S. Nguyen (2004). A strategic model for dynamic traffic assignment. *Networks and Spatial Economics 4*, 291–315.

Henderson, J. M. (2004). A planning model for optimizing locations of changeable message signs. Master's thesis, University of Waterloo.

Higle, J. L. and S. W. Wallace (2003). Sensitivity analysis and uncertainty in linear programming. *Interfaces 33*, 53–60.

Hogema, J. H., A. R. A. Vanderhorst, and P. J. Bakker (1994). Evaluation of the 16 fog-signaling system with respect to driving behavior (evaluatie van het a 16 mistsignaleringssysteem in terman van het rijgedrag). Technical Report TNO-TM 1994 C-48, TNO Technische Menskunde.

Huynh, N., Y.-C. Chiu, and H. S. Mahmassani (2003). Finding near-optimal locations for variable message signs for real-time network traffic management. *Transportation Research Record 1856*, 34–53.

Ibrahim, A. T. and F. L. Hall (1994). Effect of adverse weather conditions on speed-flow-occupancy relationships. *Transportation Research Record 1457*, 184–191.

Iida, Y. and H. Wakabayashi (1989). An approximation method of terminal reliability of a road network using partial minimal path and cut set. *Proceedings of the 5th WCTR 37*, 367–380.

Jayakrishnan, R., W. T. Tsai, J. N. Prashker, and S. Rajadhyaksha (1994). A faster path-based algorithm for traffic assignment. *Transportation Research Record 1443*, 75–83.

Jones, B., L. Janssen, and F. Mannering (1991). Analysis of the frequency and duration of freeway accidents in Seattle. *Accident Analysis and Prevention 23*, 239–255.

Kahneman, D. and A. Tversky (1979). Prospect theory: an analysis of decision under risk. *Econometrica 47*, 263–291.

Karlaftis, M. G. and I. Golias (2002). Effects of road geometry and traffic volumes on rural roadway accident rates. *Accident Analysis and Prevention 34*, 357–365.

Karoonsoontawong, A. (2006). *Robustness Approach to the Integrated Network Design Problem, Signal Optimization, and Dynamic Traffic Assignment Problem*. Ph. D. thesis, The University of Texas at Austin.

Kaysi, I. and N. H. Ali (2000). Analytical modeling of driver-guidance schemes with flow variability considerations. *Transportation Research Record 1717*, 55–65.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983). Optimization by simulated annealing. *Science 220*(4598), 671–680.

Kobayashi, K. and M. Do (2005). The informational impacts of congestion tolls upon route traffic demands. *Transportation Research Part A 39*, 651–670.

Kockelman, K. M. (2003). *Handbook of Transportation Engineering*, Chapter 12. McGraw Hill.

Labbé, M., P. Marcotte, and G. Savard (1998). A bilevel model of taxation and its application to optimal highway pricing. *Management Science 44*(12), 1608–1622.

Lam, W. H. K. and M. L. Tam (2007). Risk analysis of traffic and revenue forecasts for road investment projects. *Journal of Infrastructure Systems 4*(1), 19–27.

Lamm, R., E. M. Choueiri, and T. Mailaender (1990). Comparison of operating speeds on dry and wet pavements of two-lane rural highways. *Transportation Research Record 1280*, 199–207.

Larsson, T. and M. Patriksson (1992). Simplicial decomposition with disaggregated representation for the traffic assignment problem. *Transportation Science 26*, 4–17.

Lawphongpanich, S. and D. W. Hearn (1984). Simplicial decomposition of the asymmetric traffic assignment problem. *Transportation Research Part B 18*(2), 123–133.

Lindley, J. (1987). Urban freeway congestion: quantification of the problem and effectiveness of potential solutions. *ITE Journal 57*, 27–32.

Lindsey, R. (2008). Cost recovery from congestion tolls with stochastic capacity and demand. Presented at the Third International Conference on Funding Transport Infrastructure, Paris, France.

Liu, H. X., X. He, and W. Recker (2007). Estimation of the time-dependency of values of travel time and its reliability from loop detector data. *Transportation Research Part B 41*(4), 448–461.

Liu, H. X., W. Recker, and A. Chen (2004). Uncovering the contribution of travel time reliability to dynamic route choice using real-time loop data. *Transportation Research Part A 38*(6), 435–453.

Lium, A.-G., T. G. Crainic, and S. W. Wallace (2009). A study of demand stochasticity in service network design. *Transportation Science 43*(2), 144–157.

Lo, H. K. and Y. K. Tung (2003). Network with degradable links: capacity analysis and design. *Transportation Research Part B 37*, 345–363.

Lomax, T., D. Schrank, S. Turner, and R. Margiotta (2003). Selecting travel reliability measures. Technical report, Texas Transportation Institute and Cambridge Systematics.

Loui, R. P. (1983). Optimal paths in graphs with stochastic or multidimensional weights. *Communications of the ACM 26*(9), 670–676.

Mandel, B., M. Gaudry, and W. Rothengatter (1994). Linear or nonlinear utility functions in logit models? The impact on German high-speed rail demand forecasts. *Transportation Research Part B 28*, 91–101.

Marcotte, P. (1983). Optimization with continuous control parameters. *Transportation Science 17*, 181–187.

Marcotte, P. and S. Nguyen (1998). *Hyperpath Formulations of Traffic Assignment*, pp. 175–199. Kluwer Academic Publishers.

Marcotte, P., S. Nguyen, and A. Schoeb (2004). A strategic flow model of traffic assignment in static capacitated networks. *Operations Research 52*(2), 191–212.

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance 7*(1), 77–91.

McCord, M. and O. Villoria (1987). Nonlinear utility in time and cost of trips: disaggregate results from an ordinal methodology. *Transportation Research Record 1138*, 8–17.

Miller-Hooks, E. D. (2001). Adaptive least-expected time paths in stochastic, time-varying transportation and data networks. *Networks 37*(1), 35–52.

Miller-Hooks, E. D. and H. S. Mahmassani (2000). Least expected time paths in stochastic, time-varying transportation networks. *Transportation Science 34*(2), 198–215.

Mirchandani, P. and H. Soroush (1987). Generalized traffic equilibrium with probabilistic travel times and perceptions. *Transportation Science 21*(3), 133–152.

Mohring, H. and M. Harwitz (1962). *Highway Benefits: An Analytical Framework*. Evanston, IL: Northwestern University Press.

Montemanni, R. and L. Gambardella (2004). An exact algorithm for the robust shortest path problem with interval data. *Computers and Operations Research 31*(10), 1667–1680.

Morales, J. (1986). Analytical procedures for estimating freeway traffic congestion. *Public Roads 50*, 55–61.

Murthy, I. and S. Sarkar (1996). A relaxation-based pruning technique for a class of stochastic shortest path problems. *Transportation Science 30*(3), 220–236.

Nagae, T. and T. Akamatsu (2005). Dynamic revenue management of a toll road project under transportation demand uncertainty. *Networks and Spatial Economics 6*, 345–357.

Nam, D. and F. Mannering (2000). An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A 35*, 85–102.

Nguyen, S. and S. Pallottino (1989). *Hyperpaths and Shortest Hyperpaths*, pp. 258–271. Berlin, Germany: Springer-Verlag.

Nie, Y. and Y. Fan (2006). Arriving-on-time problem: discrete algorithm that ensures convergence. *Transportation Research Record 1964*, 193–200.

Nie, Y. and X. Wu (2009). Shortest path problem considering on-time arrival probability. *Transportation Research Part B 43*(6), 597–613.

Noland, R. B. and J. W. Polak (2002). Travel time variability: a review of theoretical and empirical issues. *Transport Reviews 22*(1), 39–54.

Ozbay, K. and P. Kachroo (1999). *Incident Management in Intelligent Transportation Systems*. Boston, MA: Artech House.

Ozbay, K. and N. Noyan (2006). Estimation of incident clearance times using Bayesian networks approach. *Accident Analysis and Prevention 38*, 542–555.

Pigou, A. C. (1920). *The Economics of Welfare*. London: Macmillan and Co.

Pinjari, A. and C. Bhat (2006). On the nonlinearity of response to level of service variables in travel mode choice models. *Transportation Research Record 1977*, 67–74.

Polychronopoulos, G. H. and J. N. Tsitsiklis (1996). Stochastic shortest path problems with recourse. *Networks 27*(2), 133–143.

Preparata, F. P. and M. I. Shamos (1985). *Computational Geometry.* New York, NY: Springer-Verlag.

Pretolani, D. (2000). A directed hypergraph model for random time dependent shortest paths. *European Journal of Operational Research 123*, 315–324.

Provan, J. S. (2003). A polynomial-time algorithm to find shortest paths with recourse. *Networks 41*(2), 115–125.

Sen, S., R. Pillai, S. Joshi, and A. K. Rathi (2001). A mean-variance model for route guidance in advanced traveler information systems. *Transportation Science 35*(1), 37–49.

Shao, H., W. H. K. Lam, and M. L. Lam (2005). A reliability based stochastic traffic assignment model for network with multiple user classes under uncertainty in demand. *Networks and Spatial Economics 6*, 169–172.

Sivakumar, R. A. and R. Batta (1994). The variance-constrained shortest path problem. *Transportation Science 28*(4), 309–316.

Skabardonis, A., P. Varaiya, and K. F. Petty (2003). Measuring recurrent and nonrecurrent traffic congestion. *Transportation Research Record 1856*, 118–124.

Small, K. A., C. Winston, and J. Yan (2005). Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica 73*(4), 1367–1382.

Smith, K. and B. Smith (2002). Forecasting the clearance time of freeway accidents. Technical Report STL-2001-012, Center for Transportation Studies, University of Virginia.

Smith, M. J. (1979). The existence, uniqueness, and stability of traffic equilibria. *Transportation Research Part B 15B*, 443–451.

Smith, M. J. (1983a). An algorithm for solving asymmetric equilibrium problems with a continuous cost-flow function. *Transportation Research Part B 17B*(5), 365–371.

Smith, M. J. (1983b). The existence and calculation of traffic equilibria. *Transportation Research Part B 17B*(4), 291–303.

Smith, T. E., E. A. Eriksson, and P. O. Lindberg (1994). Existence of optimal tolls under conditions of stochastic user-equilibria. In B. Johansson and L. G. Mattsson (Eds.), *Road Pricing: Theory, Empirical Assessment and Policy*, pp. 65–87. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Sternbach, M. C. (2001). Markowitz revisited: mean-variance models in financial portfolio analysis. *SIAM Review 43*(1), 31–85.

Supernak, J., C. Kaschade, and D. Steffey (2003). Dynamic value pricing on I-15 in San Diego: Impact on travel time and its reliability. *Transportation Research Record 1839*, 45–54.

Transportation Research Board (2000). *Highway Capacity Manual*. Washington, DC: National Research Council.

Tsaavchidis, M. (2000). Aggregate analysis of driver response to collective route guidance and implications for system control. In *Proceedings of the Tenth International Conference on Road Transport Information and Control (Conf. Publ. No. 472)*, pp. 17–21.

Tsaggouris, G. and C. Zaroliagis (2004). Non-additive shortest paths. In S. Albers and T. Radzhik (Eds.), *Proceedings of the 12th Annual European Symposium on Algorithms.*

Ukkusuri, S. and S. T. Waller (2004). An expression for long term demand uncertainty in the robust traffic assignment problem. Presented at the Fifth Triennial Symposium on Transportation Analysis (TRISTAN V), Le Gosier, Guadeloupe.

Ukkusuri, S. and S. T. Waller (2006). Single-point approximations for traffic equilibrium problem under uncertain demand. Presented at the 85th Annual Meeting of the Transportation Research Board, Washington, DC.

Ukkusuri, S. V. S. K. (2005). *Accounting for Uncertainty, Robustness, and Online Information in Transportation Networks.* Ph. D. thesis, The University of Texas at Austin.

Unnikrishnan, A. (2008). *Equilibrium Models Accounting for Uncertainty and Information Provision in Transportation Networks.* Ph. D. thesis, The University of Texas at Austin.

Valdez-Diaz, D. M., Y. C. Chiu, and H. S. Mahmassani (2001). Optimal time-dependent variable message sign diversion strategy. Presented at the 79th Annual Meeting of the Transportation Research Board, Washington, DC.

Verhoef, E. T., R. H. M. Emmerink, P. Nijkamp, and P. Rietveld (1996). Information provision, flat and fine congestion tolling and the efficiency of road usage. *Regional Science and Urban Economics 26*, 505–529.

Wallace, S. W. (2000). Decision making under uncertainty: is sensitivity analysis of any use? *Operations Research 48*(1), 20–25.

Waller, S. T., K. Kockelman, D. Sun, S. Boyles, D.-Y. Lin, M. Ng, S. Seraj, M. Tassabehji, V. Valsaraj, and X. Wang (2008). Archiving, sharing, and quantifying reliability of traffic data. Technical report, Texas Department of Transportation Report 0-5686-1.

Waller, S. T., J. L. Schofer, and A. K. Ziliaskopoulos (2001). Evaluation with traffic assignment under demand uncertainty. *Transportation Research Record 1771*, 69–74.

Waller, S. T. and A. K. Ziliaskopoulos (2002). On the online shortest path problem with limited arc cost dependencies. *Networks 40*(4), 216–227.

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

Wirasinghe, S. C. (1978). Determination of traffic delays from shock-wave analysis. *Transportation Research 12*, 343–348.

Yang, H. (1999a). Evaluating the benefits of a combined route guidance and road pricing system in a traffic network with recurrent congestion. *Transportation 26*, 299–322.

Yang, H. (1999b). System optimum, stochastic user equilibrium, and optimal link tolls. *Transportation Science 33*(4), 354–360.

Yang, H. and M. G. H. Bell (1998). Models and algorithms for road network design: a review and some new developments. *Transport Reviews 18*(3), 257–278.

Yu, G. and J. Yang (1998). On the robust shortest path problem. *Computers and Operations Research 25*(6), 457–468.

Vita

Stephen David Boyles was born in Honolulu, Hawaii on October 19, 1982 to Jean Hom Boyles (*née* Jean Lee Hom) and David Franklyn Boyles. After graduating from Rogers High School in Puyallup, Washington in 2000, he enrolled at the University of Washington in Seattle, Washington, where he receieved Bachelor of Science degrees in civil engineering and mathematics, graduating *magna cum laude* in 2004. In August 2004, he entered The Graduate School at The University of Texas at Austin under the supervision of Dr. S. Travis Waller. He received a Master of Science degree in civil engineering in May 2006, with an emphasis in transportation engineering, and immediately began doctoral studies under the same advisor. After graduation, he moved to the University of Wyoming as an assistant professor.

Permanent address: 1858 N 9th St Apt 3
Laramie, Wyoming 82072-2167

This dissertation was typeset with LaTeX[†] by the author.

[†] LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.