INFORMATION BASED MANAGEMENT OF TRANSPORT NETWORKS: NEW MODELS, ALGORITHMS AND INSIGHTS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Zhen Tan

May 2018

INFORMATION BASED MANAGEMENT OF TRANSPORT NETWORKS:
NEW MODELS, ALGORITHMS AND INSIGHTS

Zhen Tan, Ph.D.
Cornell University 2018

We studied several important management problems in modern transportation systems based on proper use of various types of information in different ways. We focused on two main research questions: 1) How to design smart information schemes for decentralizing better (e.g., more efficient, stable and sustainable) flow patterns on transportation networks (Chapter 1 & 2); and 2) How to utilize information and system data fully and efficiently for better (e.g., closer to optimal and more cost-effective) centralized decision making (Chapter 3 & 4). Specifically, we have explored the following four dimensions (each corresponds to one chapter).

1. Strategic information scheme design for enforcing optimal flow on traffic networks with minimal tolls. We explored how disclosing flexible new information can help reduce the toll intensity needed for decentralizing a Nash equilibrium on a general traffic network that minimizes certain system-level cost. We formulated the “Minimal Toll Information Design Problem” (MTIDP) and designed efficient algorithms for finding near-optimal solutions to the problem. Numerical examples are used to reveal insights of MTIDP and validate the effectiveness of the proposed solution algorithms.

2. Remedy of the negative effect of inaccurate travel time estimate on dynamic routing using additional endogenous information feedback. We proposed to provide en-route real-time traffic-sensitive pollution information to

drivers for suppressing traffic oscillations caused by delay in travel time reporting. Theoretical analysis (based on a novel queueing model), numerical examples, and simulation experiments for simple traffic networks are adopted to demonstrate the potential traffic stabilizing benefit of this new information.

3. Utilization of system data and demand forecast for real-time control of complex human-centered infrastructure systems. We used multi-access managed lanes systems as an illustrative application. With the available measurement of traffic condition and demand forecast, we developed a hybrid model predictive control based dynamic pricing algorithm using origin-destination specific tolls. Through proper formulation of system models and practical constraints, the proposed control model can be implemented efficiently in real-time.

4. Value of information and optimal learning in solving large scale network optimization problems with uncertainty. We looked at the challenging second-best network pricing problem (SNPP) with stochastic demand. We designed Bayesian learning model for the problem and tailored linear belief-based Knowledge Gradient sampling policy to SNPP. Experiment on a benchmark network with more than a million candidate solutions showed superior performance of our approach to the benchmark heuristic.

We have proposed novel methodology and generated new insights in each dimension, concrete examples are involved. Our goal is to provide useful references, practical solutions and new thinking for existent nontrivial problems and emerging challenges in traffic management under this information age and unprecedented demand for efficient and sustainable urban mobility.

We have two notes:

1) Each chapter uses independent notation, the notation table provided in one chapter (if any) is only valid within that chapter.

2) Chapter 3 and Chapter 4 are based on the following two published papers:

Z. Tan, H. Gao (2018). “Hybrid model predictive control based dynamic pricing of managed lanes with multiple accesses.” Transportation Research Part B: Methodological, 112: 113-131.

Z. Tan, H. Gao (2016). “Bayesian Ranking and Selection Model for Second-Best Network Pricing Problem.” Proceedings of the 2016 Winter Simulation Conference, 2487-2498.

BIOGRAPHICAL SKETCH

Zhen Tan is currently a 5th year Ph.D. student at Cornell University. He received a B.S. in Transportation Engineering from Tongji University in July 2009 and worked as an engineer for one year. He received an M.S. in Civil Engineering with minors in Automation and Environmental Engineering from Zhejiang University in March 2013, and expects to receive a Ph.D. in Civil and Environmental Engineering with minors in Applied Mathematics and Operations Research from Cornell University in May 2018.

ACKNOWLEDGEMENTS

Firstly, I would like to extend my sincere gratitude to my advisor Prof. H. Oliver Gao for his continuous support and guidance of my Ph.D. study and dissertation research, for his kindness, encouragement, and great vision, which helped me enhance my skills in thinking, writing and communication.

I also would like to thank the other members of my dissertation committee: Prof. Peter Frazier, Prof. Eilyan Bitar, and Prof. Sid Banerjee, for their insightful comments, suggestions, and questions that helped me reflect my research from different perspectives and explore various novel solutions.

Besides, I want to give my thanks to Prof. Eva Tardos and Prof. Jamol Pender for their advice and comments on part of my dissertation, which helped me improve the manuscript.

Finally, I would like to thank my family: my parents and my brother for their loving care, understanding and spiritual support during my Ph.D. years and throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
MINIMAL NETWORK PRICING WITH STRATEGIC INFORMATION PROVISION

In the first chapter, we explore how disclosing new (and presumably imperceivable) travel-related information can help reduce (as much as possible) the toll intensity needed for decentralizing a system optimal (SO) flow on a traffic network. The information considered in our model is very general: it can be either endogenous or exogenous and can be in various types such as available local link information in the routing map, travel time variability estimate, and health risk due to air pollution exposure in traffic, etc.. Although there has been an increasing number of studies discussing how providing extra information to users may affect traffic flow distribution on the network and the resulting externalities, the analysis on the full potential or benefit new information disclosure can bring about is still lacking. Therefore, we go one step further to discuss optimal information scheme design models and algorithms.

Specifically, we consider a flexible setting where one or more of the following dimensions can be included in the information scheme design: 1) selection of user groups; 2) selection of origin-destination pairs, and 3) selection of the links (where the relevant information are to be disclosed). The performance of a certain information scheme is measured by the minimal toll intensity (an increasing function of link tolls) needed for enforcing an SO flow. The SO flow is defined as the link flow pattern that minimizes the total system cost such as time delay or emissions. We derive necessary conditions for toll intensity reduction by proving new information regarding all the links of the network to all the users and sufficient conditions for toll intensity reduction by partial information disclosure (e.g., only on a subset of links). We then formulate a general minimum toll information design problem (MTIDP) to find the information distribution schemes that needs the lowest possible toll intensity for decentralizing an SO flow. Two practical algorithms are proposed for solving the MTIDP. Numerical examples are used to verify the effectiveness of the proposed solution methods. Our results imply that the potential of using information intervention in network control depends significantly on how much flexibility is allowed in the information design as well as the combination of different system parameters such as user preferences and cost functions on the links. Useful insights and caveats are also generated by solving a number of representative MTIDP instances. Therefore the proposed optimization model and solution algorithms can be very useful tools for designing effective and robust information schemes for proactive traffic management.

1.1 Introduction

Congestion pricing has been studied extensively in theory (e.g., [44, 129]) and implemented as an regulation of road transport externalities in practice (e.g., [29, 52]). While some projects are going smoothly, many proposals faltered due to social and political resistance (e.g., [64]) though proved effective in easing gridlock by analytical or simulation studies. High toll charge is a main factor that limits the applicability of congestion pricing projects due to concerns such as the equity effect [38, 48, 64]. Without intelligent mechanism design and proper revenue redistribution, pricing policy can be regressive [75], burdening low-income drivers and even keeping them from promising workplaces [38]. To avoid paying the toll, violation behavior are often reported (e.g., [25]), causing difficulty in regulation enforcement. Furthermore, considering other externalities of traffic congestion such as emissions and air pollution, the welfare maximizing toll will be much higher [20, 29] and less affordable. Therefore, more “soft”, fair, acceptable and sustainable traffic management policies are desirable for proactive congestion mitigation in addition to regulatory tolls. For example, by leveraging big data and operations research techniques, we can improve the information distribution schemes by taking different travelers' behavior and preferences into account, so that a system optimal flow distribution can be enforced with only a marginal toll charges. In this study, we investigate a general model for such a flexible information scheme design problem on traffic networks targeted to heterogeneous users. The objective of our information design problem is minimizing system-level cost (e.g., total delay and emissions) with the least intensive tolls.

Intelligent transportation systems (ITS) technologies, GPS-enabled on-board devices and smart phones as well as growing computing power make data collection, processing and information release more convenient. Traffic conditions and travel time forecasts has been provided in many cities for traffic management [123]. Development in advanced traveler general information systems (ATGIS) enables disclosing of more comprehensive and different types of travel-related information such as fuel consumption and green house gas emissions [105] via fixed or portable media [80]. Furthermore, advancement in sensing and estimation technologies makes it feasible to forecast and thus disclose information on various kinds of externalities/travel characteristics such as high-resolution air pollution and exposure levels [21, 30, 33]. The explicit provision of the new information would affect travelers' decision making due to more directly perceived factors such as reliability, safety and health risk (e.g., due to exposure to air pollution), which were usually underestimated or even neglected.

There is an emerging stream of research about how new information can affect traffic on the network. For example, it was proved that providing different "information sets" (defined as the available links of the network) to different user groups can lead to less efficient Nash flow compared to the case where all the users have the full information about the network [7]. Recent study [119] also shows that the increased penetration of routing apps can have negative externalities as more users have knowledge on low-capacity local road links and the traffic volume increases dramatically on those residential streets, which causes bad impact on local community and costs government millions to steer away the traffic. Numerical experiments of realistic network by [105] showed that providing travelers with fuel and emissions costs via ATGIS could cut the total delay on the network by 1% to 17% or even increase it by 1%, depending on the perceived cost prior to the disclosure of these new information if any. This interesting observation reflects that the impact of new information provision on system output can be very complicated considering how users react to and value the new information. Another important complexity comes from users can have different risk attitude towards travel time variability, even though most users are risk-averse, their relative valuation of travel time variability compared to the expected travel time can vary notably [88]. The model and argument of work [100] indicate a potential of efficiency loss by providing travel time variability information to the selfish routers who are risk-averse and underestimate or ignore the travel time variability.

The above-mentioned studies ([7, 100, 105, 119]) and other related work focused on analyzing and revealing interesting effect of information provision on traffic efficiency on the network, but did not discuss how different types of information can be strategically provided (possibly with other traffic management measures) for optimizing system performance. Furthermore, although there are previous studies (e.g., [78, 112]) discussed the minimum toll problem (in the sense of toll revenue), none of them consider how information can be used to reduce the toll for an SO flow. Therefore, we go one step further to discuss the optimal information design problem with the present of pricing. Specifically, we formulate the problem of determining where and to whom such information should be disclosed on the network that can minimize the link-based anonymous toll charges needed for enforcing a system optimal (SO) flow pattern. Flexible practical constraints on the spatial distribution and targets of the new information can be incorporated in our model. We define the SO objective function to be a system performance measure (e.g., total traffic delay or emissions on the network) that is monotone in link flows. We consider a finite number of heterogeneous user groups with fixed demand and assume the new information brings some extra perceived cost which is link-additive (such as variance of travel time, en-route exposure to air pollution).

The rest of this chapter is organized as follows. In Section 2, we present the mathematical models for network routing under new information provision and link-based pricing and characterize the feasible toll set that can realize a system optimal (SO) Nash flow given any information scheme. Section 3 raises the question of whether toll intensity needed for enforcing an SO network flow can be reduced using strategic information design, both necessary condition for effective full information scheme and sufficient condition for partial information scheme are discussed. In Section 4, we systematically study this toll reduction potential by formulating the "minimum-toll information design problem"

(MTIDP) and propose two efficient approximation algorithms for solving the MTIDP. Numerical examples are presented in Section 5 and we draw conclusions and discuss future extensions in Section 6.

1.2 Mathematical Formulation

We have $G = (N,\ A)$ be a directed transportation network defined by a set N of nodes and a set A of directed links. Each link $a \in A$ has an associated average travel time t_a and an piece of extra information e_a. Suppose each link a can also have a nonnegative toll charge τ_a. A path p is a sequence of links which connect a sequence of nodes following the link directions. Let W be the set of origin-destination (OD) pairs, each is a pair of distinct nodes in N. Let P_w be the set of all paths connecting OD pair $w \in W$, Each OD pair w has a fixed demand $d_w > 0$ to be routed through the network. This traffic network problem in the computer science community is also referred to as the multi-commodity flow problem (e.g., [44]) where each "commodity" $w \in W$ goes from the source-destination pair defined by the OD pair w.

1.2.1 Models for link travel time and extra information

Monotonic univariate link volume-delay function is commonly used in traffic assignment models to account for not only link but also intersection delay (which can be considerable in urban networks) (e.g., [109]). Thus we assume

Assumption 1.1 *The average travel time t_a is a function of the link flow variable x_a,* *denoted by function $t_a(x_a)$ and is continuous and strictly increasing in x_a.*

Now we introduce some quantity $e_a \in \mathbb{R}$ associated with each link a. For simplicity and demonstration of the key idea, we assume here that e_a is imperceivable unless it is disclosed or reported explicitly. For example, e_a can be an exogenous quantity such as the exposure to some air pollutant that is dominated by background emissions or some environmental conditions independent of the flow [80]. It can also represents the exposure to some air pollutant that is very sensitive to traffic condition, then intuitively e_a should depend on traffic intensity not only on link a but also on all the links near link a [29, 30]. Also, e_a can be negative, which represents some type of "discount" or "credit" by traveling on link a (i.e., availability of new ITS facility that reduces en-route speed fluctuations). Thus for generality, we express each extra link-specific quantity e_a as a function of the flow distribution on the entire network:

$$e_a = e_a(x), \tag{1.1}$$

where column vector $x = \{x_a, \ \forall a \in A\}$ is the link-flow vector. Note that we can define $e_a(x) \equiv e_a$ to be some proper constant to represent special type of information, such as the availability of the link in routing [7, 119] (in which case we can set $e_a \equiv \infty$ or a very large number to "hide" link a from or to not suggest link a on the routing map). We need a key assumption regarding e_a on different links when we evaluate the same quantity for a certain path.

Assumption 1.2 *The quantity e_a is link-additive, i.e., this quantity for one path p (denoted as e^p) is equal to $\sum_{a \in p} e_a$.*

The total travel demand consists of groups with different socio-economic characteristics and travel purposes that govern their behavior parameters. Let

M be the set of groups of users, so each group is indexed by $m \in M$. Let $\alpha_m \in \mathbb{R}_{++}$ and $\beta_m \in \mathbb{R}$ be the average value of time (VOT) and value of extra information (VOE), respectively, for user group m. Notice that in order to model general cases of user valuation of the new information, there is no restriction on the sign of β_m in our model. For easy of analysis we assume no two groups have the same VOT and sort the user groups in the increasing order of α_m, i.e., $0 < \alpha_1 < \alpha_2 < ... < \alpha_{|M|}$. We define $d_w^m \geq 0$ as the demand of group m for OD pair w, so

$$\sum_{m \in M} d_w^m = d_w. \tag{1.2}$$

We define a binary vector $\delta \in \{0,1\}^{|A||W||M|}$ which has $\delta_a^{w,m} = 1$ ($a \in A,\ w \in W,\ m \in M$) if the quantity e_a is provided on link a for user group m of OD pair w, and $\delta_a^{w,m} = 0$ otherwise. We call δ the "info-selection vector". We denote $\hat{e}_a^{w,m}$ the extra information regarding link a observed by user group m on OD pair w, hence we have

$$\hat{e}_a^{w,m} = \delta_a^{w,m} e_a. \tag{1.3}$$

On each link a, there is also a toll charge $\tau_a \geq 0$, which is anonymous (i.e., every user on link a needs to pay the same amount τ_a regardless of what user group she belongs to), anonymous tolls are realistic and easy to enforce [130].

Therefore, on each OD pair $w \in W$ for each user group $m \in M$, each path $p \in P_w$ is associated with a travel time t^p, a perceived extra information $\hat{e}_m^p$, and an anonymous toll charge $\tau^p \geq 0$. By construction, these quantities are

$$t^p = \sum_{a \in p} t_a(x_a),\ \hat{e}_m^p = \sum_{a \in p} \delta_a^{w,m} e_a(x),\ \tau^p = \sum_{a \in p} \tau_a,\ p \in P_w,\ w \in W, \tag{1.4}$$

the second equality is because of (1.3) and Assumption 1.2.

1.2.2 User equilibrium (UE) and system optimal (SO) flows

Suppose each user chooses a path that minimizes his own travel cost that includes the cost of travel time, and possibly the extra perceived cost associated with the extra information if any, as well as toll charges if any. Since we assumed the quantity e_a is imperceivable and hard to estimate, we assume that

Assumption 1.3 *Users take into account the extra cost associated with quantity e_a on link a only when they are informed of this value on link a (i.e., the perceived information for link a is $\hat{e}_a = e_a$).*

Under provision of the extra information according to info-selection vector δ, users' perceived travel cost includes factors other than travel delay, so it fits to a conventional concept "generalized cost" used in literature, and is usually measured by unified monetary cost (e.g., [93, 130]). Here in our problem both travel time t_a and the quantity e_a are transferred into equivalent amount of money using the multiplicative parameters VOT and VOE, respectively, when evaluating the generalized cost of users. We denote $g_a^{w,m}$ and $\hat{g}_p^m$ as the generalized cost for users of group m of OD pair w traveling on link a and on path p.We assume a linear perceived cost function for users. Then we have

$$g_a^{w,m} = \alpha_m t_a + \beta_m \delta_a^{w,m} e_a + \tau_a, \ a \in A, \ w \in W, \ m \in M; \tag{1.5}$$

$$\hat{g}_p^m = \alpha_m t^p + \beta_m \hat{e}_m^p + \tau^p, \ p \in P_w, \ w \in W, \ m \in M \tag{1.6}$$

We define the column vector that contains all the $g_a^{w,m}$ or $\hat{g}_p^m$ terms

$$g = \{g_a^{w,m}, w \in W, m \in M, a \in A\} = \{g_1^{1,1} ... g_1^{1,|M|} ... g_1^{|W|,1} ... g_1^{|W|,|M|} ... g_{|A|}^{|W|,1} ... g_{|A|}^{|W|,|M|}\}; \tag{1.7}$$

$$\hat{g} = \{\hat{g}_p^m, m \in M, p \in P_w, w \in W\} = \{\hat{g}_1^1 ... \hat{g}_1^{|M|} ... \hat{g}_{|P_1|}^1 ... \hat{g}_{|P_1|}^{|M|} ... \hat{g}_{|P_{|W|}|}^1 ... \hat{g}_{|P_{|W|}|}^{|M|}\}. \tag{1.8}$$

We also define the OD-group-link flow vector y and group-path flow vector f

$$\begin{aligned} y &= \{y_a^{w,m}, w \in W, m \in M, a \in A\} = \{y_1^{1,1} ... y_1^{1,|M|} ... y_1^{|W|,1} ... y_1^{|W|,|M|} ... y_{|A|}^{|W|,1} ... y_{|A|}^{|W|,|M|}\}; \quad (1.9) \\ f &= \{f_p^m, m \in M, p \in P_w, w \in W\} = \{\hat{f}_1^1 ... \hat{f}_1^{|M|} ... \hat{f}_{|P_1|}^1 ... \hat{f}_{|P_1|}^{|M|} ... \hat{f}_{|P_{|W|}|}^1 ... \hat{f}_{|P_{|W|}|}^{|M|}\}, \quad (1.10) \end{aligned}$$

where $y_a^{w,m}$ is the flow of user group m of OD pair w on link a, and f_p^m is the flow of user group m on path p. The generalized cost vector g (or $\hat{g}$) can be expressed as a function of the OD-group-link flow vector y (or the group-path flow vector f) as $g(y)$ (or $\hat{g}(f)$) since it is a function of the link flow x.

We call the group-path flow vector f feasible if it lies in

$$\mathcal{F} := \left\{ f \in \mathbb{R}_+^{|M| \sum_{w \in W} |P_w|} : \sum_{p \in P_w} f_p^m = d_w^m, \ \forall m \in M, \ w \in W \right\}. \tag{1.11}$$

We call the OD-group-link flow vector y feasible if it lies in

$$\mathcal{Y} := \left\{ y \in \mathbb{R}^{|W||M||A|} : \ \exists f \in \mathcal{F}, \ \text{s.t.}\ y_a^{w,m} = \sum_{p \in P_w : a \in p} f_p^m, \ \forall a \in A, \ m \in M \right\}. \tag{1.12}$$

Finally we call the link-flow vector x feasible if it lies in

$$\mathcal{X} := \left\{ x \in \mathbb{R}^{|A|} : \exists y \in \mathcal{Y}, \ \text{s.t.}\ x_a = \sum_{m \in M} y_a^m, \ \forall a \in A \right\}. \tag{1.13}$$

Note that $\mathcal{F} \subseteq \mathbb{R}_+^{|M| \sum_{w \in W} |P_w|}$, $\mathcal{Y} \subseteq \mathbb{R}_+^{|W||M||A|}$ and $\mathcal{X} \subseteq \mathbb{R}_+^{|A|}$ are all closed and convex.

With the generalized cost described in (1.5), we define f^{UE} as a user equilibrium (UE) group-path flow pattern (a.k.a. Nash flow [44]) if under f^{UE} no user has incentive to unilaterally switch his path [109], i.e.,

$$\hat{g}_p^m(f^{UE}) \begin{cases} = \gamma_w^m, & \text{if} f_p^m > 0 \\ \geq \gamma_w^m, & \text{if} f_p^m = 0 \end{cases} \quad \forall p \in P_w, \ w \in W, \tag{1.14}$$

where γ_w^m represents a common quantity for all the paths that are used by user group m between OD pair w under the UE flow. In other words, all utilized paths by a certain group of users on a certain O/D pair have equal and minimal generalized cost. We also define y^{UE} as the OD-group-link UE flow. The UE flow y^{UE} or f^{UE} can be characterized by the following variational inequalities.

Theorem 1.1 *A OD-group-link flow $y^{UE} \in \mathcal{Y}$ (a group-path flow $f^{UE} \in \mathcal{F}$) is a user equilibrium if only if*

$$g(y^{UE})^{\mathrm{T}}(y - y^{UE}) \geq 0,\ \forall y \in \mathcal{Y} \quad \left(\hat{g}(f^{UE})^{\mathrm{T}}(f - f^{UE}) \geq 0,\ \forall f \in \mathcal{F}\right) \tag{1.15}$$

Condition (1.15) can be defined in either group-path flow or OD-group-link flow.The proof is omitted as this is a natural extension of the classical result in network equilibrium analysis, see, e.g., [93, 114].

Remark 1.1 *In general there is no equivalent mathematical programing formulation for the UE flow problem with more than one cost components that are associated with arbitrary weights. Hence the above variational inequality characterization serves for the basis of the algorithms that find the equilibrium numerically. In addition, we have a very general function $e_a(x)$ to represent the possibly endogenous extra quantity, so the user equilibrium flow pattern is not necessarily unique [71, 93]. However, when int he following three special cases, we claim that the UE flow is unique: 1) There is only one group of user, $|M| = 1$; 2) There are more than one groups of users but the ratios α_m/β_m ($m \in M$) are the same; 3) there is no extra information (i.e., $\delta = \mathbf{0}$, where $\mathbf{0}$ is a vector of zeros) or the quantity e_a is exogenous (i.e., $e_a(x) = e_a$) $\forall a \in A$, This can be seen as a straightforward corollary of the Theorem in Section 2.1 of [71]. In particular, we will utilize the special case 3) in our discussion later.*

The focus of our study is not computing the user equilibrium. In stead, we are interested in how the extra information in addition to toll charges can enable an Nash flow pattern that coincides with some desired target flow pattern on the network using as low tolls as possible. The target flow pattern can be any system optimal (SO) flow under which some centralized network performance measure is optimized. For example, the operator may want to minimize (over the peak period), the total delay on the network; the total emission on the network; or the linear combinations of the two or more objectives. Here we consider the SO flow pattern in terms of the minimization of a general type of objective functions that satisfy the property below.

Assumption 1.4 *The SO performance measure Φ can be expressed as a function of the link-flow vector x, $\Phi(x)$. In addition, $\partial\Phi/\partial x_a > 0,\ \forall a \in A$.*

A large class of practical objectives commonly used in the literature satisfies the above property. For example, one can consider the following objectives: 1) Total travel time on the network (SO-TT), which is a traditional and important network performance measure ([44, 109, 130]; 2) Total emissions on the network (SO-TEM), which is an important overall environmental objective [9, 66]; 3) A linear combination of the above two objectives (SO-TT&TEM) that gives flexibilities in trading off total travel time and total emissions [9]. All the objectives are evaluated over the period within which the OD demand $\{d_w\}$ is modeled. The minimization problems are

$$\mathrm{SO-TT}: \qquad \min_{x\in\mathcal{X}} \sum_{a\in A} x_a t_a(x_a); \tag{1.16}$$

$$\mathrm{SO-TEM}: \qquad \min_{x\in\mathcal{X}} \sum_{a\in A} l_a x_a r_a(x_a); \tag{1.17}$$

$$\mathrm{SO-TT\&TEM}: \qquad \min_{x\in\mathcal{X}} \omega^{TT} \sum_{a\in A} x_a t_a(x_a) + \omega^{TEM} \sum_{a\in A} x_a l_a r_a(x_a) \tag{1.18}$$

where $\omega^{TT}, \omega^{TEM} > 0$ are the weights associated with the total travel time and total emissions, respectively, in the objective SO-TT&TEM; r_a is the vehicle emission rate per distance on link a, which is a increasing and convex function of link flow a; l_a is the length of link a.

It can be verified that the objectives in (1.16) - (1.18) are all convex in link flow x. The feasible region $\mathcal{X}$ is closed and convex. Hence the optimal solutions to (1.16) - (1.18) can be computed very efficiently using any convex optimization techniques. Also note that when t_a and r_a are strictly convex, then the SO link-flow is unique, but it may correspond to multiple optimal OD-group-link flows y and group-path flows f.

Related to Remark 1.1 we made above, here we make an observation based on the definition of SO flow and extra information.

Proposition 1.1 *Given a SO link flow x^*, and any info-selection vector δ, if we fix the extra information for group $m \in M$ of OD pair $w \in W$ regarding link $a \in A$ as*

$$\hat{e}_a^{w,m} = \delta_a^{w,m} e_a^* = \delta_a^{w,m} e_a(x^*), \tag{1.19}$$

then the UE flow is unique.

Proof: Note that under δ and e_a^* the general cost for a user group $m \in M$ on link a is $g_a^{w,m} = c_a^m + \alpha_m t_a(x_a)$, where $c_a^m = \beta_m \delta_a^{w,m} e_a(x^*) + \tau_a$ is a group-specific constant term. Then the uniqueness of the UE flow follows readily from the Theorem in Section 2.1 of [71] as we noted in Remark 1.1 since (1.19) represents a special case of "exogenous" information. ■

Proposition 1.1 is fundamental for our following discussion of enforcing an SO flow by tolls and information. Specifically, by using link-based tolls together with new information about e_a's whose values are determined by a given SO flow, we are able decentralize an SO flow as a Nash equilibrium.

1.2.3 System optimal tolls under extra information

Suppose we want to generate a Nash flow x^{UE} that coincides with some SO flow pattern x^* (both in terms of link flow here) by setting proper nonnegative toll charges on the links, which is required to be anonymous, i.e., each user of on the same link pays the same toll specified for this link regardless of which group she belongs to. Let the column vector $\tau = \{\tau_a\} \in \mathbb{R}^{|A|}$, we call τ SO-flow enabling if it can reproduce a SO Nash flow $x^{UE} = x^*$, and denote $\mathcal{T}$ as the set of all such eligible toll vectors.

It is known that pricing according to the total marginal social cost on each link (thus the link toll is anonymous) leads to the SO flow when 1) the users are homogeneous; 2) the non-toll cost considered by users and that defined for the SO objective consist of the same quantities with the same weights. For example, under our problem setting, if users have homogeneous VOT ($\alpha^1 = \alpha$) and VOE ($\beta^1 = \beta$) (m=1), the information design is link-based and independent of ODs (subscripts m and w are not needed for δ), and the SO-objective is

$$\sum_{a \in A} \alpha t_a(x_a) + \beta \delta_a e_a(x), \tag{1.20}$$

then putting a toll on each link a that equals the total marginal social cost (assuming both $t_a(\cdot)$ and $e_a(\cdot)$ are differentiable)

$$\alpha t_a(x_a^*) \frac{dt_a(x_a)}{dx_a}\bigg|_{x_a = x_a^*} + \beta \delta_a e_a(x_a^*) \frac{\partial e_a(x)}{\partial x_a}\bigg|_{x = x^*}$$

results in a UE flow that minimizes the above objective. This can be easily verified by the first order condition of the minimization problem with objective (1.20) and feasible region $\mathcal{X}$. Details are omitted here.

Nevertheless, if users are heterogeneous or the SO objective is different from (1.20) or information design is OD-dependent, such marginal cost tolling can not work in general. Instead, it turns out that $\mathcal{T}$ can be characterized by the optimal dual solutions of a linear programing (LP) problem. Suppose x^* is an SO-link-flow that minimizes function $\Phi(x)$, and $t_a^* = t_a(x_a^*)$ and $e_a^* = e_a(x^*)$, $\forall a \in A$. We formulate the LP problem

$$
\begin{aligned}
\min_{f,\, z} \quad & \sum_{w \in W} \sum_{p \in P_w} \sum_{m \in M} f_p^m \left(\alpha_m t^{*p} + \beta_m \hat{e}_m^{*p}\right) \\
\text{s.t.} \quad & z_a + \sum_{m \in M} \sum_{w \in W} \sum_{p \in P_w : a \in p} f_p^m = x_a^*, \ \forall a \in A \\
& \sum_{p \in P_w} f_p^m = d_w^m, \ \forall m \in M, \ w \in W \\
& f_p^m \geq 0, \ \forall p \in P_w, \ w \in W, \ m \in M \\
& z_a \geq 0, \ \forall a \in A,
\end{aligned}
\tag{1.21}
$$

where $z = \{z_a, \ \forall a \in A\}$ contains the slack variables. The dual of problem (1.21) is

$$
\begin{aligned}
\max_{\tau, \gamma} \quad & \sum_{w \in W} \sum_{m \in M} d_w^m \gamma_w^m - \sum_{a \in A} x_a^* \tau_a \\
\text{s.t.} \quad & \gamma_w^m - \sum_{a \in A : a \in p} \tau_a \leq \alpha_m t^{*p} + \beta_m \hat{e}_m^{*p}, \ \forall p \in P_w, \ w \in W, \ m \in M \\
& \tau_a \geq 0, \ \forall a \in A.
\end{aligned}
\tag{1.22}
$$

where the column vectors $\gamma = \{\gamma_w^m, \ \forall m \in M, \ w \in W\}$ and $\tau = \{\tau_a, \ \forall a \in A\}$ contain all the dual variables. Here τ_a is actually the negate of the multiplier associated with the corresponding constraint in (1.21), thus the coefficients of each τ_a in the objective and the first constraint of (1.22) are negative instead of positive. This makes the meaning of τ_a more relevant (the toll on link a).

Note that the primal problem (1.21) is feasible because there must exist a group-path flow vector $f \in \mathcal{F}$ such that $\sum_{m \in M} \sum_{p : a \in p} f_p^m = x_a^*$ since we have the SO link flow solution $x^* \in \mathcal{X}$. In addition, the primal objective value is lower bounded by 0 since $f_p^m \geq 0$ for feasibility. This implies both the primal problem (1.21) and its dual (1.22) have an optimal solution. Suppose $(\tilde{f}, \ \tilde{z})$ is an optimal solution to the primal problem and $(\tilde{\tau},\ \tilde{\gamma})$ is an optimal solution to the dual problem. Thus by strong duality, we know the two optimal values are equal, in addition, $(\tilde{f},\ \tilde{z})$ and $(\tilde{\tau},\ \tilde{\gamma})$ are complementary slack, i.e.,

$$\begin{cases} \tilde{f}_p^m \left(\tilde{\tau}^p + \alpha_m t^{*p} + \beta_m \hat{e}_m^{p^*} - \gamma_w^m\right) = 0,\ \forall p \in P_w,\ w \in W,\ m \in M; \\ \tilde{z}_a \tilde{\tau}_a = 0,\ \forall a \in A, \end{cases} \tag{1.23}$$

where $\tilde{\tau}^p = \sum_{a \in p} \tilde{\tau}_a$.Based on the first line in (1.23) in addition to the primal and dual feasibility constraints $\tilde{f}_p^m \geq 0$, $\tilde{\tau}^p + \alpha_m t^{*p} + \beta_m \hat{e}_m^{*p} \geq \tilde{\gamma}_w^m$, we know that the generalized cost

$$\hat{g}_p^m(\tilde{f}) \begin{cases} = \tilde{\gamma}_w^m, & \text{if} \tilde{f}_p^m > 0 \\ \geq \tilde{\gamma}_w^m, & \text{if} \tilde{f}_p^m = 0 \end{cases} \quad \forall p \in P_w,\ w \in W. \tag{1.24}$$

This is exactly the UE condition in the form of (1.14), i.e., for each OD pair w only the path(s) with minimal generalized cost γ_w^m are used by user group m. Hence $\tilde{f}$ corresponds to a UE flow pattern.

Now let $\tilde{x}$ be the link-flow corresponding to $\tilde{f}$, we show that $\tilde{x} = x^*$ by contradiction. We know that $\tilde{x} \in \mathcal{X}$ has its entry $\tilde{x}_a \leq x_a^*$, $\forall a \in A$. If there exits $a \in A$ such that $\tilde{x}_a < x_a^*$, then by Assumption 1.4, we know that $\Phi(\tilde{x}) < \Phi(x^*)$, which implies that x^* is not an SO link-flow. In other words, in the optimal solution $(\tilde{f},\ \tilde{z})$ to the primal problem (1.21), we must have that $\tilde{z}_a = 0$, $\forall a \in A$ (i.e., $\sum_{m \in M} \sum_{w \in W} \sum_{p \in P_w} \tilde{f}_p^m = \tilde{x}_a = x_a^*$).

Therefore, we have demonstrated that a toll vector specified by any optimal solution to the dual problem (1.22) enforces an SO-flow pattern on the network if the SO-objective function satisfies Assumption 1.4. This result is a natural extension of the one proved in [44] (where only delay cost is considered by the users) to the setting where there are cost components other than travel delay (such as the perception of air pollution exposure as we consider here). Study

[130] also used similar argument to characterize the SO-flow enabling tolls considering only delay cost for the special SO objective – total delay on the network, i.e., the one defined in (1.16). Notice that the converse is also true: any toll vector τ that can result in a SO-flow pattern must form an optimal solution to the dual problem (1.22) together with the minimum generalized cost vector γ. We summarize these observations in the following theorem.

Theorem 1.2 *Consider the following constraints in the link-toll vector τ and the minimum group-OD generalized cost vector γ*

$$\begin{cases} \sum_{a\in p}\tau_a \geq \gamma_w^m - \sum_{a\in p}\alpha_m t_a(x_a^*) + \beta_m \delta_a^{w,m} e_a(x^*), \ \forall p \in P_w, \ w \in W, \ m \in M; \\ \sum_{a\in A}\tau_a x_a^* = \sum_{w\in W}\sum_{m\in M} d_w^m \gamma_w^m - \sum_{m\in M}\sum_{a\in A}\tilde{y}_a^m \left[\alpha_m t_a(x_a^*) + \beta_m e_a(x^*)\sum_{w\in W}\delta_a^{w,m}\right]; \\ \tau_a \geq 0, \ \forall a \in A, \end{cases} \tag{1.25}$$

where x^* is any minimizer of $\Phi(x)$ over $\mathcal{X}$, and $\tilde{y}$ is the OD-group-link flow vector that corresponds to any optimal solution $\tilde{f}$ of the problem (1.21). Define the set

$$\mathcal{T} := \{\tau : \ (\tau, \ \gamma) \text{ satisfies } (1.25)\}. \tag{1.26}$$

Then we have that 1) $\mathcal{T}$ is nonempty; 2) If Assumption 1.4 holds, a link-toll vector τ results in a UE flow pattern x^* (which minimizes the SO-objective function Φ) if only if $\tau \in \mathcal{T}$.

Proof: The inequalities in (1.25) are equivalent to the dual feasibility constraints in (1.22), and the equality in (1.25) means the primal and the dual optimal values are the same by strong duality by noting that $\sum_{w\in W}\sum_{p\in P_w}\sum_{m\in M} f_p^m\left(\alpha_m t^{*p} + \beta_m \hat{e}^{*p}\right) = \sum_{m\in M}\sum_{a\in A} y_a^m \left(\alpha_m t_a^* + \beta_m e_a^* \sum_{w\in W}\delta_a^{w,m}\right)$. Thus the "if" part is proved earlier.

Now we show the converse. Suppose a toll vector $\tilde{\tau}$ results in a UE flow x^* that minimizes $\Phi(x)$, $\tilde{f} \in \mathcal{F}$ is a resultant group-path flow vector, and $\tilde{\gamma}$ is the group-OD generalized cost vector under $\tilde{\tau}$. Then the UE condition (1.24) is satisfied, which implies (by complementary slackness) that $(\tilde{f}, \tilde{\tau}, \tilde{\gamma})$ is an optimal primal-dual solution to the problem (1.21). Hence $\tilde{\tau} \in \mathcal{T}$. ■

We know by (1.24) that for each γ_w^m, there must be at least one inequality in the first line of (1.25) is tight. Thus given an optimal primal solution $\tilde{f}$ and due to strong duality, we can actually characterize the set $\mathcal{T}$ in (1.26) explicitly by simply eliminating variables γ_w^m using those equalities in the first line of (1.25) and express the rest strict inequalities in terms of only variables τ_a.

As we will discuss later, making the extra information $\hat{e}_a$ associated with extra quantity e_a perceivable has a potential of cutting the toll charges that are needed for achieving the SO flow pattern, for which Theorem 1.2 serves as a starting point.

1.3 Cutting the Tolls for the SO Flow using Extra Information

Our focus in this study is to reduce the minimum toll intensity needed to realize the SO-flow pattern using new information. To this end, we define the toll intensity measure J and assume it to have the following property.

Assumption 1.5 *The toll intensity measure J can be expressed as a convex function of the link-toll vector τ, $J(\tau)$. In addition, $\partial J/\partial \tau_a > 0, \text{ at } \tau_a > 0 \; \forall a \in A$ and $J(\mathbf{0}) = 0$.*

The toll intensity $J(\tau)$ that satisfies Assumption 1.5 can have many forms for various policy needs.For example, we can consider minimizing the weighted sum of the tolls on the links (with strictly positive weights); or a convex quadratic function of the toll vector, i.e.,

$$J(\tau) = \omega^{\mathrm{T}}\tau;\ J(\tau) = \tau^{\mathrm{T}}\Lambda\tau,$$

where ω is a wight vector whose entry a, $\omega_a > 0$ represents how much we want to restrict the toll on link a, Λ is a positive diagonal matrix used to penalize the corresponding link tolls.

As explained earlier, our goal is to minimize the toll intensity needed for an SO-flow, i.e., minimize $J(\tau)$ over $\mathcal{T}$. Note that the feasible toll set $\mathcal{T}$ as defined in (1.26) actually depends on how the extra information is disclosed (encoded by the vector δ), so it is a function of δ, $\mathcal{T}(\delta)$. Thus, we call a toll "minimal" under δ if it minimizes J over $\mathcal{T}(\delta)$. Then given an info-selection vector, δ, the minimization problem is

$$\min_{\tau \in \mathcal{T}(\delta)} J(\tau). \tag{1.27}$$

Suppose under the SO-link flow x^*, the travel time on each path p is t^{*p}. For each OD pair $w \in W$, we sort the paths such that $t^{*p_1^w} \leq t^{*p_2^w} \leq ... \leq t^{*p_{|P_w|}^w}$. Under no provision of the new information, we suppose $f(\mathbf{0})$ reproduces the SO flow x^*. We then define a mapping $Q_w^0(p)$ for used paths $p \in P_w$ under $f(\mathbf{0})$ by the following: if p has a distinct delay among all used paths in P_w, we simply map it to itself; if multiple used paths have the same delay (then they must have the same toll cost due to UE condition), we map them to a common element q. We denote the range of this mapping as $\mathcal{Q}_w^0$ whose elements have the consistent ordering with the original ordering of the paths in P_w. For example, if $|P_w| = 4$ and $t^{p_1^w} < t^{p_2^w} = t^{p_3^w} < t^{p_4^w}$, and all four paths are used under $f(\mathbf{0})$, then we have

$\boldsymbol{Q}_w^0 = \{q_1^w, q_2^w, q_3^w\}$, where p_1^w is mapped to q_1^w, p_4^w is mapped to q_3^w, and both p_2^w and p_3^w are mapped to q_2^w. We also define two quantities for each user group m

$$\bar{q}_w^0(m) : \quad = \quad \text{argmax}\left\{q \in \boldsymbol{Q}_w^0 : \exists p \in P_w \text{ s.t. } f(\mathbf{0})_p^m > 0,\ Q_w^0(p) = q\right\} \tag{1.28}$$

$$\underline{q}_w^0(m) : \quad = \quad \text{argmin}\left\{q \in \boldsymbol{Q}_w^0 : \exists p \in P_w \text{ s.t. } f(\mathbf{0})_p^m > 0,\ Q_w^0(p) = q\right\} \tag{1.29}$$

We have a basic observation regarding the equilibrium behavior of the users under the original scenario without new information: any path chosen by a user with a higher VOT is at least as fast as any path chosen by a user with a lower VOT. Formally, we have (WOLG we assume $d_w^m > 0,\ \forall m \in M$)

Lemma 1.1 *Under flow $f(\mathbf{0})$, for each OD pair $w \in W$, each user group $m \in M$ exactly uses paths $p \in P_w$ such that $Q_w^0(p) = \underline{q}_w^0(m), \underline{q}_w^0(m) + 1, ..., \bar{q}_w^0(m)$. I.e., the image of the paths used by each user group under mapping Q_w^0 is connected. In addition, $\bar{q}_w^0(m) \leq \underline{q}_w^0(m-1),\ \forall m = 2, ..., |M|$.*

Proof: Let $\tau(\mathbf{0})$ be any link toll vector that results in flow $f(\mathbf{0})$ when $\delta = \mathbf{0}$. Show by contradiction. Suppose under $f(\mathbf{0})$, among all the used paths in P_w, user group m uses paths $p_1,\ p_3$ but not path p_2 with $Q_w^0(p_1) < Q_w^0(p_2) < Q_w^0(p_3)$. So there must be some other user group $m' \neq m$ uses path p_2, which means

$$\text{if } m' < m, \qquad \text{then } \tau(\mathbf{0})^{p_2} - \tau(\mathbf{0})^{p_3} \leq \alpha_{m'}(t^{*p_3} - t^{*p_2}) < \alpha_m(t^{*p_3} - t^{*p_2});$$

$$\text{if } m' > m, \qquad \text{then } \tau(\mathbf{0})^{p_1} - \tau(\mathbf{0})^{p_2} \geq \alpha_{m'}(t^{*p_2} - t^{*p_1}) > \alpha_m(t^{*p_2} - t^{*p_1}),$$

in both cases, group m has a incentive to use path p_2, so it is a contradiction.

Now we show $\bar{q}_w^0(m) \leq \underline{q}_w^0(m-1),\ \forall m = 2, ..., |M|$ also by contradiction. Suppose $\bar{q}_w^0(m) > \underline{q}_w^0(m-1)$ for some $m = 2, ..., |M|$, then by definition (1.28)-(1.29) we know $t^{*\bar{q}^w,m} > t^{*\underline{q}^w(m-1)}$. So based on the connected image argument above, this implies that there exist two paths $p, p' \in P_w$ with $t^{*p} > t^{*p'}$ such that both groups $m-1$ and m use both paths p and p'. So by UE condition,

$$\alpha_{m-1}(t^{*p} - t^{*p'}) = \tau(\mathbf{0})^{p'} - \tau(\mathbf{0})^{p} = \alpha_m(t^{*p} - t^{*p'}),$$

which is a conflict since $\alpha_{m-1} < \alpha_m$. ■

1.3.1 Necessary conditions for effective full information

As a first attempt, the operator can simply disclose the new information on every link of the network to every user (i.e., set $\delta = \mathbf{1}$, where $\mathbf{1}$ is a vector of ones). However, it is nontrivial to conclude if this can result in a reduction in the toll intensity for realizing a SO flow pattern or not, i.e., if the optimal objective value (1.27) decreases or not:

$$\min_{\tau \in \mathcal{T}(\mathbf{1})} J(\tau) < \min_{\tau \in \mathcal{T}(\mathbf{0})} J(\tau),$$

where$\mathcal{T}(\mathbf{0})$ and $\mathcal{T}(\mathbf{1})$ are, respectively, the SO-flow enabling toll set defined in (1.26) when there is no extra information ($\delta = \mathbf{0}$) and extra information is posted on every link ($\delta = \mathbf{1}$). Obtaining $\mathcal{T}(\mathbf{0})$ or $\mathcal{T}(\mathbf{1})$ needs to solve the primal LP problem (1.21), hence we have to solve four LPs in order to check the above inequality. However, we also want to explore under what conditions this toll reduction may happen. Let's look at a basic example.

Consider a simple network that has two nodes A and B, and only two links connect A to B, shown in Figure 1.1 (where functions $t_a(\cdot)$ and $e_a(\cdot)$ are given). Suppose there are two groups of users $m = 1, 2$ with travel demand $d^1 = d^2 = 10$ and VOT and VOE parameters as $\alpha_1 = 1$, $\alpha_2 = 2$, $\beta_1 = 0.2$, $\beta_2 = 0.1$. The SO link flows in terms of total delay (1.16) are $x_1^* = 8.33$, $x_2^* = 11.67$. Two possible scenarios of functional forms $e_a(\cdot)$ are compared. On the left plot $e_a(\cdot)$ are the same on the two links $e_a = 0.1x_a t_a$, $(a = 1, 2)$, while on the right plot $e_1(\cdot)$ is the same with that in the left plot, but $e_2 = 0.3t_2$ is just a linear function of travel time t_2. Table 1.1 shows the travel times t_a and quantities e_a under the SO solution, the minimal SO-flow enabling tolls τ^* (minimizer of $J(\tau)$ over $\mathcal{T}$), the total toll charged (optimal value $J^* = J(\tau^*) = \sum_{a=1,2} \tau_a^* x_a^*$) as well as the group-path flows under the minimal toll solution.

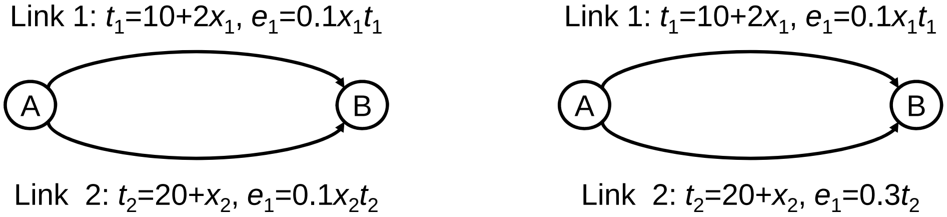

Link 1: t_1=10+2x_1, e_1=0.1x_1t_1
A
B
Link 2: t_2=20+x_2, e_1=0.1x_2t_2
Link 1: t_1=10+2x_1, e_1=0.1x_1t_1
A
B
Link 2: t_2=20+x_2, e_1=0.3t_2

Figure 1.1: An example (left: same function $e_a(\cdot)$ on two links; right: different functions $e_a(\cdot)$ on two links).

Table 1.1: Outputs of the simple network example

Scenario	No new info.	With new info. (left)	With new info. (right)
t_1^*, t_2^*	26.67, 31.67	26.67, 31.67	26.67, 31.67
e_1^*, e_2^*	/, /	22.22, 36.94	22.22, 8
f_1^1, f_2^1	0, 10	0, 10	0, 10
f_1^2, f_2^2	8.33, 1.67	8.33, 1.67	8.33, 1.67
τ_1^*, τ_2^*	10, 0	11.47, 0	8.58, 0
J^*	83.33	95.60	71.48

We can see that if functions $e_a(\cdot)$ are the same on two links (left plot), the minimum total toll charged is higher than that when no extra information is provided, while if functions $e_a(\cdot)$ are different on two links (right plot), the minimum total toll charged reduces compared to the case when there is no extra information. Note that the group-path flows are unchanged in both scenarios that have the extra information present compared to the original scenario where the extra information is not disclosed. This means that the optimal solutions to the primal LP (1.21) are the same under all the three scenarios. In addition, it can be checked that in all three scenarios the minimal toll is only applied to link 1, the purpose of which is making group 2 indifferent to which link of the two to choose (i.e., $g_2^1 = g_2^2$), while group 1 always prefers link 2. This explains why in the left plot the toll increases while in the right one the toll decreases: in the example of the left plot, $t_1^* < t_2^*$ and $e_1^* < e_2^*$ which actually enlarges the difference of the non-toll costs between the two links for any group, hence a higher toll on link 1 is needed to make group 2 still has same generalized cost of joining either link as compared to the case when the extra information is not disclosed. In contrast, in the example of the right plot, $t_1^* < t_2^*$ but $e_1^* > e_2^*$, since the differences $t_2^* - t_1^*$ and $e_1^* - e_2^*$ have same magnitude but β_m is much smaller than α_m, $(m = 1, 2)$, hence for both groups the non-toll cost of the two links become closer, therefore a lower toll is needed on link 1 to make $g_2^1 = g_2^2$. We formalize a related result of the above simple example.

Proposition 1.2 *For a simple network (parallel links connecting one OD pair), if* $\min_{\tau \in \mathcal{T}(\mathbf{1})} J(\tau) < \min_{\tau \in \mathcal{T}(\mathbf{0})} J(\tau)$, *then we have either* $\exists a,\ a' \in A,\ m \in M$ *such that* $t_a^* > t_{a'}^*$ *but* $\beta_m e_a^* < \beta_m e_{a'}^*$, *or* $\exists m,\ m' \in M,\ a \in A$ *such that* $\alpha_m > \alpha_{m'}$ *but* $\beta_m e_a^* < \beta_{m'} e_{a'}^*$, *or both.*

Proof: For a simple network, each path p is just a link a between the same OD pair, so the analysis is the same in terms of paths. We show by contradiction. Suppose any two paths $p, p' \in P$ with $t^{*p} \geq t^{*p'}$ have $e^{*p} \geq e^{*p'}$ and any two groups $m, m' \in M$ with $\alpha_m > \alpha_{m'}$ have $\beta_m \geq \beta_{m'}$. Let $(f(\mathbf{1}), \tau(\mathbf{1}), \gamma(\mathbf{1}))$ be an optimal primal-dual solutions to (1.21) under $\delta = \mathbf{1}$, and $\tau(\mathbf{1})$ is also a minimizer of $J(\tau)$ over $\mathcal{T}(\mathbf{1})$. Consider any fixed $m \in M$ and those paths $p \in P$ that have $f(\mathbf{1})_p^m > 0$.

suppose there are $K_m \geq 1$ such paths, we sort them as $p_1^m, ..., p_{K_m}^m$ in the increasing order of their travel times. Then under $\delta = \mathbf{1}$ the corresponding inequalities in (1.25) for these paths are tight by complementary slackness, i.e.,

$$\gamma(\mathbf{1})^m \quad = \quad \tau(\mathbf{1})^p + \alpha_m t^{*p} + \beta_m e^{*p}, \ \forall p \in \left\{p_1^m, ..., p_{K_m}^m\right\}, \tag{1.30}$$

it follows that

$$\tau(\mathbf{1})^{p_k^m} - \tau(\mathbf{1})^{p_{k+1}^m} = \alpha_m(t^{*p_{k+1}^m} - t^{*p_k^m}) + \beta_m(e^{*p_{k+1}^m} - e^{*p_k^m}), \tag{1.31}$$

hence $\tau(\mathbf{1})^{p_1^m}, ..., \tau(\mathbf{1})^{p_{K_m-1}^m}$ are determined by $\tau(\mathbf{1})^{p_{K_m}^m}$.

Now because any two paths $p,\ p'$ with $t^{*p} \geq t^{*p'}$ have $\beta_m e^{*p} \geq \beta_m e^{*p'}$ for any user group m, and any two groups $m,\ m'$ with $\alpha_m > \alpha_{m'}$ have $\beta_m e^{*p} \geq \beta_{m'} e^{*p}$ for any path p, it is straightforward to have a similar result to Lemma 1.1 (where $\delta = \mathbf{0}$) for $\delta = \mathbf{1}$: a higher indexed user group uses paths that are at least equally fast and two adjacent user groups use at most one same path. Then we have $\tau(\mathbf{1})^{p_{K_1}^1} \leq ... \leq \tau(\mathbf{1})^{p_1^1} \leq \tau(\mathbf{1})^{p_{K_2}^2} \leq ... \leq \tau(\mathbf{1})^{p_1^2} \leq ... \leq \tau(\mathbf{1})^{p_{K_{|M|}}^{|M|}} \leq ... \leq \tau(\mathbf{1})^{p_1^{|M|}}$ by (1.31).

Moreover, we claim that the tolls $\tau(\mathbf{1})^p$ of all the used paths p in P can be determined from $\tau(\mathbf{1})^{p_{K_1}^1}$. This is so by considering two possible cases of any adjacent user groups m and $m-1$: 1) if $\bar{q}^1(m) = \underline{q}^1(m-1)$, then $\tau(\mathbf{1})^{p_{K_m}^m} = \tau(\mathbf{1})^{p_1^{m-1}}$; 2) if $\bar{q}^1(m) > \underline{q}^1(m-1)$, then it must be true that

$$\begin{cases} \tau(\mathbf{1})^{p_{K_m}^m} \geq \tau(\mathbf{1})^{p_1^{m-1}} + \alpha_{m-1}(t^{*p_1^{m-1}} - t^{*p_{K_m}^m}) + \beta_{m-1}(e^{*p_1^{m-1}} - e^{*p_{K_m}^m}); \\ \tau(\mathbf{1})^{p_{K_m}^m} \leq \tau(\mathbf{1})^{p_1^{m-1}} + \alpha_m(t^{*p_1^{m-1}} - t^{*p_{K_m}^m}) + \beta_m(e^{*p_1^{m-1}} - e^{*p_{K_m}^m}), \end{cases} \tag{1.32}$$

where we define a mapping Q_w^1 for $\delta = \mathbf{1}$ in a similar manner as Q_w^0 for $\delta = \mathbf{0}$, and $\bar{q}^1(m),\ \underline{q}^1(m)$ are defined for $\delta = \mathbf{1}$ in a similar manner as $\bar{q}^0(m),\ \underline{q}^0(m)$ for $\delta = \mathbf{0}$. Hence it follows that the first of the above two inequalities is tight because $\alpha_m(t^{*p_1^{m-1}} - t^{*p_{K_m}^m}) > \alpha_{m-1}(t^{*p_1^{m-1}} - t^{*p_{K_m}^m}) \geq 0, \beta_m(e^{*p_1^{m-1}} - e^{*p_{K_m}^m}) \geq \beta_{m-1}(e^{*p_1^{m-1}} - e^{*p_{K_m}^m}) \geq 0,$

and $\tau(\mathbf{1})$ minimizes J over $\mathcal{T}(\mathbf{1})$ (i.e., increasing $\tau(\mathbf{1})^{p^m_{K_m}}$ makes objective J larger), and this equality is achievable since the network is simple (link toll is just path toll). We also deduce that $\tau(\mathbf{1})^{p^1_{K_1}} = 0$ since $\tau(\mathbf{1})$ minimizes J over $\tau(\mathbf{1})$.

Then based on $\tau(\mathbf{1})$ we construct a toll vector $\tau(\mathbf{0})$ as follows:

$$\tau(\mathbf{0})^{p^m_{K_m-j}} = \tau(\mathbf{1})^{p^m_{K_m-j}} - \beta_m(e^{*P^m_{K_m-j+1}} - e^{*P^m_{K_m-j}}),\ j = 1, ..., K_m - 1,\ \forall m$$
$$\tau(\mathbf{0})^{p^m_{K_m}} = \begin{cases} \tau(\mathbf{1})^{p^{m-1}_1} - \beta_{m-1}(e^{*P^{m-1}_1} - e^{*P^m_{K_m}}), & \text{if } \bar{q}^1(m) = \underline{q}^1(m-1) \\ \tau(\mathbf{0})^{p^{m-1}_1} + \alpha_{m-1}(t^{*P^{m-1}_1} - t^{*P^m_{K_m}}), & \text{if } \bar{q}^1(m) > \underline{q}^1(m-1) \\ 0, & \text{if } m = 1. \end{cases} \tag{1.33}$$

Thus $\tau(\mathbf{0})^p \leq \tau(\mathbf{1})^p,\ \forall p \in P$ (equivalently, $\tau(\mathbf{0})_a \leq \tau(\mathbf{1})_a,\ \forall a \in A$). In addition, it can be seen that $f(\mathbf{1})$ is also a UE flow under $\tau(\mathbf{0})$ with $\delta = \mathbf{0}$, and since $f(\mathbf{1})$ is a SO flow, thus $\tau(\mathbf{0}) \in \mathcal{T}(\mathbf{0})$ by the optimality condition of the problem (1.21). Therefore, we deduce by Assumption 1.5 that

$$\min_{\tau\in\mathcal{T}(\mathbf{0})} J(\tau) \leq J(\tau(\mathbf{0})) \leq J(\tau(\mathbf{1})) = \min_{\tau\in\mathcal{T}(\mathbf{1})} J(\tau),$$

which is contradictory to $\min_{\tau\in\mathcal{T}(\mathbf{1})} J(\tau) < \min_{\tau\in\mathcal{T}(\mathbf{0})} J(\tau)$. ■

The above result has some policy implications. For example, in managed lane systems where the tolled lanes and the general purpose lanes run in parallel, minimizing total delay may need higher tolls if information about the air pollution exposure (which is proportional to the travel time and pollution concentration [21]) is disclosed on both types of lanes users have similar valuation to it (and more pollution exposure means no higher utility, i.e., $\beta_m \leq 0\ \forall m \in M$). This is because the air pollutant concentrations are very similar on two type of lanes and thus the ranking of e_a are the same with t_a and disclosing e_a enlarges the perceived cost difference between two types of lanes. However, if the paths consists of multiple transport modes, it may be effective in lowering the tolls by providing such extra information for all the modes. For example, it is likely that driving on the ground takes shorter but has higher exposure to air pollution, while mass transit via elevated viaduct takes longer but is cleaner.

The above basic result implies that "full" provision of the information can have negative benefit in reducing the toll intensity. Instead, disclosing extra information only on a subset of links or a subset of users may be more effective in toll intensity reduction. However, allowing flexible "partial" provision of information can significantly increase the possible number of information schemes, which makes the design problem nontrivial and we will discuss this in the remaining of the chapter.

1.3.2 Sufficient conditions for effective partial information

Recall the example in Figure 1.1, if in the left plot we only provide the extra information on link 1 to both users, then the optimal toll becomes $\tau_1^* = 7.78$. $\tau_2^* = 0$ and results an even lower total toll charged of 64.81, since the cost of e_1 serves just as an extra "price" on link 1 that reduces the gap between the non-toll cost on two links. Essentially, via strategic provision of the extra information "partially" on the network, we may re-balance the perceived travel cost which results in a lower total toll charge needed to regulate the traffic towards a SO flow pattern. The intuition behind the effectiveness in toll reduction by partial information is rooted in the necessary condition of toll reduction by full information discussed in the previous subsection. For example, we have the following straightforward observation.

Lemma 1.2 *There exists an info-selection vector δ such that $t^{*p} < t^{*p'}$ but $\beta_m \hat{e}_m^{*p} >$*

$\beta_m \hat{e}_m^{*p'}$ for some user group $m \in M$, some paths $p,\ p' \in P_w$ and some OD pair $w \in W$. And there exits an info-selection vector δ such that $\alpha_m > \alpha_{m'}$ but $\beta_m \hat{e}_m^{*p} < \beta_{m'} \hat{e}_{m'}^{*p}$ for some user groups $m,\ m' \in M$, some path $p \in P_w$ and some OD pair $w \in W$.

Proof: Pick two paths $p \neq p' \in P_w$ and some $w \in W$ with $t^{*p} < t^{*p'}$, and a link a such that $a \in p,\ a \notin p'$ and $e_a^* \neq 0$. Suppose $\beta_m \neq 0$ for some user group $m \in M$. If $\beta_m e_a^* > 0$, we can set $\delta_a^{w,m} = 1,\ \delta_b^{w,m} = 0, \forall b \neq a$, then we have $\beta_m e_a^* = \beta_m \hat{e}_m^{*p} > \beta_m \hat{e}_m^{*p'} = 0$; same argument applies to the case $\beta_m e_a^* < 0$.

Similarly, we pick two groups $m \neq m' \in M$ with $\alpha_m > \alpha_{m'}$, and $\beta_{m'} \neq 0$. Suppose $e*_a \neq 0$ on some link $a \in p$ for some path $p \in P_w$ and some $w \in W$. If $\beta_{m'} e*_a > 0$, we can set $\delta_b^{w,m} = 0\ \forall b \in p$ and $\delta_a^{w,m'} = 1,\ \delta_b^{w,m'} = 0\ \forall b \in p$ with $b \neq a$, then we have $0 = \beta_m \hat{e}_m^{*p} < \beta_{m'} \hat{e}_{m'}^{*p} = \beta_{m'} e*_a > 0$; same argument applies to the case $\beta_{m'} e*_a < 0$. ■

Our focus on the rest of this section is to come up with some practical rules of partial information design that ensure positive toll intensity reduction. We first construct a general sufficient condition that ensures positive toll intensity reduction. This is based on the routing under the minimal SO-flow enabling tolls when there is no extra information on the network.

Theorem 1.3 *Let $\tau(\mathbf{0}) \in \arg\min_{\tau \in \mathcal{T}(\mathbf{0})} J(\tau)$, and $f(\mathbf{0})$ be an optimal solution to (1.21) under $\delta = \mathbf{0}$. For an info-selection vector δ, if there exits a toll vector $\hat{\tau}$ such that $J(\hat{\tau}) < J(\tau(\mathbf{0}))$ and $f(\mathbf{0})$ corresponds to a UE flow under $\hat{\tau}$ and δ, then $\min\limits_{\tau \in \mathcal{T}(\delta)} J(\tau) < J(\tau(\mathbf{0}))$.*

Proof: $f(\mathbf{0})$ is an optimal solution to (1.21) with $\delta = \mathbf{0}$ means that $f(\mathbf{0})$ forms a SO flow. Then since it is also a UE flow under $\hat{\tau}$ and δ, we deduce that $\hat{\tau} \in \mathcal{T}(\delta)$ by the definition of $\mathcal{T}(\delta)$. Thus $\min\limits_{\tau \in \mathcal{T}(\delta)} J(\tau) \leq J(\hat{\tau}) < J(\tau(\mathbf{0}))$. ■

By Theorem 1.3, one can attempt to cut the toll intensity by focusing on disclosing the extra information on the tolled links under $\tau(\mathbf{0})$ provided that certain conditions are satisfied. We construct some intuitive cases to focus on selecting "qualified" links out of three types of tolled links under $\tau(\mathbf{0})$: 1) links with zero flow; 2) links that are used by only group(s) of users that have the same VOE; 3) links that are used by groups of users that have at least two different VOEs. We denote $g(\mathbf{0})$ and $\gamma(\mathbf{0})$, respectively, the generalized cost, and minimum generalized cost, under $\delta = \mathbf{0}$ and $\tau(\mathbf{0})$.

Proposition 1.3 *Define $A_0 = \{a \in A : \tau(\mathbf{0})_a > 0\}$, and let M_a be the set of groups that use link a under $f(\mathbf{0})$ (where $\tau(\mathbf{0})$ and $f(\mathbf{0})$ are defined in Theorem 1.3). μ_a be the number of distinct values of β_m among the users that use link a under $f(\mathbf{0})$. Suppose link a is in either one of the sets A_1, A_2 and A_3 defined below, and we form an info-selection vector δ by setting $\delta_a^{w,m} = 1$, $\forall w \in W$, $m \in M$ (starting from $\delta = \mathbf{0}$), then we have $\min\limits_{\tau \in \mathcal{T}(\delta)} J(\tau) < J(\tau(\mathbf{0}))$.*

$$A_1 = \left\{ a \in A_0 : \begin{array}{c} \mu_a = 0,\ g(\mathbf{0})_p^m - \gamma(\mathbf{0})_w^m \geq \min\{\tau(\mathbf{0})_a, \beta_{\hat{m}_a} e_a^*\} - \beta_m e_a^*, \\ \forall p \in P_w \text{ with } a \in p,\ w \in W,\ m \in M \text{ with } \beta_m e_a^* \leq 0 \end{array} \right\}, \tag{1.34}$$

where $\hat{m}_a = \arg\min_{m \in M}\{\beta_m e_a^* : \beta_m e_a^* > 0\}$;

$$A_2 = \left\{ a \in A_0 : \begin{array}{r} \mu_a = 1, m_a \in M_a,\ \tau(\mathbf{0})_a \geq \beta_{m_a} e_a^* > 0, \\ g(\mathbf{0})_p^m - \gamma(\mathbf{0})_w^m \geq (\beta_{m_a} - \beta_m) e_a^*,\ \forall p \in P_w \text{ with } a \in p, \\ w \in W,\ m \in M \text{ with } \beta_m e_a^* < \beta_{m_a} e_a^* \end{array} \right\}; \tag{1.35}$$

$$A_3 = \left\{ a \in A_0 : \begin{array}{r} \mu_a > 1,\ \tau(\mathbf{0})_a \geq \beta_{\bar{m}_a} e_a^* > 0, \\ \sum\limits_{p \in P_w : a \in p} f(\mathbf{0})_p^m = d_w^m,\ \forall w \in W_a^m,\ m \in M_a \text{ with } \beta_m e_a^* < \beta_{\bar{m}_a} e_a^* \\ g(\mathbf{0})_p^m - \gamma(\mathbf{0})_w^m \geq (\beta_{\bar{m}_a} - \beta_m) e_a^*,\ \forall p \in P_w \text{ with } a \in p, \\ w \in W,\ m \in M \setminus M_a \text{ with } \beta_m e_a^* < \beta_{\bar{m}_a} e_a^* \end{array} \right\}, \tag{1.36}$$

where $\bar{m}_a = \arg\max_{m \in M_a} \beta_m e^*_a$ *and* $W^m_a = \{w \in W : \exists p \in P_w \text{ s.t. } a \in p,\ f^m_p > 0\}$.

Proof: The idea is to show that we can construct a toll vector $\hat{\tau}$ such that $f(\mathbf{0})$ (an optimal solution to (1.21) with $\delta = \mathbf{0}$, as defined in Theorem 1.3) is a UE group-path flow under $\hat{\tau}$ and the δ constructed by setting $\delta^{w,m}_a = 1\ \forall w, m$ for link $a \in A_0$ in either of the sets $A_i,\ i = 1, 2, 3$. Besides, $J(\hat{\tau}) \leq J(\tau(\mathbf{0}))$. Thus the result follows from Theorem 1.3. See the Appendix for detailed proof. ■

Remark 1.2 *Proposition 1.3 implies that actually starting from* $\delta = \mathbf{0}$*, we can loop over all the links* $a \in A_0$ *by checking if* $a \in A_i$ *in any order of* $i \in \{1, 2, 3\}$ *and setting* $\delta^{w,m}_a = 1\ \forall w, m$ *(if yes) and update the relevant quantities* τ*,* g *and* γ *after the application of each change of* δ*, and update* $A_i \leftarrow A_i \setminus \{a\}$ *(if* $a \in A_i$*) with the updated new variables* τ*,* g *and* γ*. Then we may achieve more toll reduction. This can be verified by simple induction as the toll intensity reduction after each update of* δ *must be strictly positive by Proposition 1.3 based on a similar statement as Theorem 1.3.*

The above three rules are only intuitive examples of how to construct an info-selection vector δ that ensures positive reduction in toll intensity for enforcing an SO flow, there may be other similar but more sophisticated rules. The main advantage of such rules is that we can actually design efficient algorithms to implement them (e.g., find the sets A_i, $i = 1, 2, 3$), this can be achieved with a reformulation of problem (1.21) with a different set of variables (see Remark 1.5 in Section 4). However, it should be emphasized that the effectiveness of such rules are highly dependent on network structure and user behavior parameters, which are problem specific. For example, we have a corollary to Proposition 1.3 when the distributions of VOTs and VOEs satisfy certain conditions.

Corollary 1.1 *If we form δ by either (i) or (ii) (when applicable) starting with $\delta = \mathbf{0}$, then we have $\min_{\tau \in \mathcal{T}(\delta)} J(\tau) \leq J(\tau(\mathbf{0}))$:*

(i) If $\beta_m = \beta\ \forall m \in M$, then for each $a \in A$, if $x_a^* = 0$, $\tau_a > 0$, $\beta e_a^* > 0$, or $x_a^* > 0$, $\tau(\mathbf{0})_a^* \geq \beta e_a^* > 0$, we set $\delta_a^{w,m} = 1\ \forall w \in W, m \in M$.

(ii) For each $a \in A$ with $x_a^* > 0$, let $W_a = \{w \in W : \exists p \in P_w \text{ s.t. } a \in p\}$ and $m_{a,w}^0 = \min\left\{m \in M : \exists p \in P_w \text{ s.t. } a \in p \text{ and } f(\mathbf{0})_p^m > 0\right\}$. If $\beta_m \geq \beta_{m+1}\ \forall m = 1, ..., |M| - 1$, then set $\delta_a^{w,m} = 1\ \forall w \in W, m \in M$ if $m_{a,w}^0 = m'\ \forall w \in W_a$, $\tau(\mathbf{0})_a^* \geq \beta_{m'} e_a^* > 0$, and link a is used by all the shortest paths between OD pairs $w \in W_a$.

Proof: We can verify that a link $a \in A$ that qualifies for condition (i) must lie in either set A_1 or A_2 defined in Proposition 1.3. And a link $a \in A$ that qualifies for condition (ii) must lie in either set A_2 or A_3. Thus setting $\delta_a^{w,m} = 1\ \forall w, m$ for a link a that qualifies either (i) or (ii) leads to $\min_{\tau \in \mathcal{T}(\delta)} J(\tau) < J(\tau(\mathbf{0}))$ by Proposition 1.3. Then we can apply the reasoning in Remark 1.2 to loop over $a \in A$ for either case (i) or (ii) and update δ starting from $\delta = \mathbf{0}$ for qualified links, which gives the result. See the Appendix for details. ■

The application of (i) in Corollary 1.1 requires that users have a common sensitivity to the extra information e_a, this may well approximate the case where the information has quite uniform perceived value especially when it is displayed in monetary units. For other types of information, it is more likely that the users have different sensitivity to e_a. And to apply (ii) when $e_a > 0$, we need that users with higher VOT have lower VOE, which is a ordering condition on the distribution of user utility function parameters. For some types of information this may be the case, for example, empirical studies (e.g., [107]) found that the air pollution information has less impact on travelers who have more urgent trips (thus value time more) (see, e.g., the second scenario of the first example in Section 5); similarly, users who are very picky about arriving on time may care less the safety implications during the trip. This assumption may not be good in other cases (e.g., users who value expected time more are likely to also care more about the reliability of the travel time [26]), and richer people (who value time more) may be also more willing to avoid higher health risk due to air pollution exposure. In addition, the number of the links that are "qualified" to set $\delta_a = 1$ heavily depends on the network structure. For some simple real networks, such as the managed lane systems on highway networks (in which higher VOT users are willing to pay the toll for faster service), it is usually the case that some link on the shortest paths is tolled. Also the minimal tolls are typically put on the links that are only used by the shortest paths of one or a few OD pairs, since intuitively, a higher toll is needed for one link if the potential demand of traveling on this link is higher. Thus cutting the toll on the qualified links according to (ii) can be effective. But for other settings, there may be very few or no qualified links (e.g., our second numerical example).

Therefore, in order to study the full potential of flexible information design in reducing the toll regulation for enforcing an optimal flow for general cases, we have to solve an optimization model. In the next section, we will formulate the mathematical programing problem for finding the info-selection vector δ that can reduce the toll intensity $J(\tau(\delta))$ for realizing an SO flow pattern on a general network with any distributions of VOTs and VOEs. This problem turns out to be very challenging to solve, we also propose two efficient algorithms to obtain near-optimal solutions.

1.4 Minimum-Toll Information Design Problem (MTIDP)

1.4.1 Problem formulation

We consider the problem of finding the optimal information scheme (encoded by info-selection vector δ) that can help achieve an SO flow patten using the least toll intensity, we call it "minimal-toll information design problem (MTIDP)". Recall that given a fixed δ, the least toll intensity needed for an SO flow can be obtained by solving problem (1.27). Since the eligible toll set $\mathcal{T}(\delta)$ defined in (1.27) can be characterized by the optimality condition of problem (1.21) under δ involving continuous variables, formally we need to solve an optimization problem on mixed-integer decision variables. The binary variables correspond to whether to provide certain type of users with the extra information related to certain links of the network or not, and the continuous variables are the primal and dual variables of problem (1.21). Specifically, by the primal and dual feasibility as well as complement slackness condition to (1.21) (i.e., optimality condition), the MTIDP can be formulated as the following nonlinear mixed integer programing (NLMIP) problem.

$$
\begin{aligned}
\min_{\delta,\,\tau,\,\gamma,\,f} \quad & J(\tau) \\
\text{s.t.} \quad & \delta \in \{0,1\}^{|W||M||A|},\ \tau \in \mathbb{R}_+^{|A|},\ \gamma \in \mathbb{R}^{|W|\times|M|},\ f \in \mathcal{F} \\
& \sum_{a\in p}(\tau_a + \alpha_m t_a^* + \beta_m \delta_a^{w,m} e_a^*) \geq \gamma_w^m,\ \forall p \in P_w,\ w \in W,\ m \in M \\
& \left(\sum_{a\in p}(\tau_a + \alpha_m t_a^* + \beta_m \delta_a^{w,m} e_a^*) - \gamma_w^m\right) f_p^m = 0,\ \forall p \in P_w,\ w \in W,\ m \in M \\
& \sum_{m\in M}\ \sum_{p\in P: a\in p} f_p^m = x_a^*,\ \forall a \in A.
\end{aligned}
\tag{1.37}
$$

Before we present these solution methods to MTIDP, we first note that it is desired to compute the feasible SO-flow enabling toll set $\mathcal{T}$ efficiently in the first place (for larger networks). In order to avoid solving an exponential sized LP (1.21) due to path enumeration, we can re-formulate the problem in link-based decision vector $\{f_a^{w,m},\ \forall w \in W,\ m \in M, a \in A\}$ (each entry is the flow of group m on link a from OD pair w) as hinted in [44]. Then we get a problem which is equivalent to (1.21)) but is instead polynomial sized:

$$\begin{aligned}
\min_{f} \quad & \sum_{m\in M}\sum_{w\in W}\sum_{a\in A} f_a^{w,m}\left(\alpha_m t_a^* + \beta_m \delta_a^{w,m} e_a^*\right) \\
\text{s.t.} \quad & \sum_{m\in M}\sum_{w\in W} f_a^{w,m} \leq x_a^*,\ \forall a \in A \qquad\qquad (1.38)\\
& \sum_{a\in A: a_s = n} f_a^{w,m} - \sum_{a\in A: a_t = n} f_a^{w,m} = \begin{cases} d_w^m, \text{ if } n = s_w \\ -d_w^m, \text{ if } n = t_w \\ 0, \text{ otherwise} \end{cases},\ \forall n \in N,\ m \in M,\ w \in W \\
& f_a^{w,m} \geq 0,\ \forall a \in A,\ m \in M,\ w \in W,
\end{aligned}$$

where $a_s,\ a_t$ denote the starting and ending nodes of link a, respectively; and $s_w,\ t_w$ denote the origin node and destination node of OD pair w, respectively. The dual of problem of (1.38) is

$$\begin{aligned}
\max_{\tau,\eta} \quad & \sum_{w\in W}\sum_{m\in M} d_w^m\left(\eta_{s_w}^{w,m} - \eta_{t_w}^{w,m}\right) - \sum_{a\in A} x_a^* \tau_a \\
\text{s.t.} \quad & \eta_{a_s}^{w,m} - \eta_{a_t}^{w,m} \leq \tau_a + \alpha_m t_a^* + \beta_m \delta_a^{w,m} e_a^*,\ \forall a \in A,\ w \in W,\ m \in M \qquad (1.39)\\
& \tau_a \geq 0,\ \forall a \in A.
\end{aligned}$$

where the column vector $\eta = \{\eta_n^{w,m},\ \forall n \in N,\ m \in M,\ w \in W\}$ contains the dual variables other than the tolls: the entry $\eta_n^{w,m}$ represents the minimal generalized cost (relative to $\eta_{t_w}^{w,m}$) a group m user would have incurred starting from node n to her destination (if she had chosen a path that leads her to n). Using similar argument, we can prove that the optimal dual solution $\tilde{\tau}$ to (1.39) results in the SO flow x^*. But now the size of the problem reduces significantly (the number of constraints is only linear in $|A|$ and $|N|$ given W and M). Accordingly, the SO-flow eligible toll set can be re-expressed as

$$\mathcal{T} := \{\tau : \ (\tau,\ \eta) \text{ is an optimal solution to problem (1.39)}\}. \tag{1.40}$$

and problem (1.37) can be reformulated as a polynomial sized NLMIP problem

$$\begin{aligned}
\min_{\delta,\ \tau,\ f,\ \eta} \quad & J(\tau) \\
\text{s.t.} \quad & \tau \in \mathbb{R}^{|A|},\ \delta \in \{0,1\}^{|W||M||A|},\ f \in \mathbb{R}^{|A||W||M|},\ \eta \in \mathbb{R}^{|N||W||M|}, \\
& \text{Constraints in (1.38) and in (1.39)} \qquad\qquad (1.41) \\
& (\eta_{a_s}^{w,m} - \eta_{a_st}^{w,m} - \tau_a - \alpha_m t_a^* - \beta_m \delta_a^{w,m} e_a^*) f_a^{w,m} = 0,\ \forall a \in A,\ w \in W,\ m \in M
\end{aligned}$$

Remark 1.3 *By Assumption 1.5, we know that if the optimal value of problem (1.41) $J(\tau^*) = 0$, we have $\tau_a^* = 0\ \forall a \in A$. I.e., we can actually enforce an SO flow purely relying on information design, which is the ideal case.*

Remark 1.4 *If we knew there exits a minimizer $(\delta^*, \tau^*, f^*, \eta^*)$ to problem (1.41) with $f^* = f(\mathbf{0})$ (the optimal solution to the primal LP (1.38) under $\delta = \delta^*$, i.e., $f(\mathbf{0})$ remains a UE flow under δ^* and τ^*, then problem (1.41) can be reduced to an linear mixed integer programing (LMIP) problem by replacing the nonlinear constraints by*

$$\begin{cases} \eta_{a_s}^{w,m} - \eta_{a_st}^{w,m} - \tau_a - \alpha_m t_a^* - \beta_m \delta_a^{w,m} e_a^* = 0, \text{ if } f(\mathbf{0})_a^{w,m} > 0 \\ \eta_{a_s}^{w,m} - \eta_{a_st}^{w,m} - \tau_a - \alpha_m t_a^* - \beta_m \delta_a^{w,m} e_a^* \le 0, \text{ if } f(\mathbf{0})_a^{w,m} = 0 \end{cases},\ \forall a \in A,\ w \in W,\ m \in M, \tag{1.42}$$

which is much easier to solve. However, since we did not know if such a minimizer exits or not, we can instead utilize this easier LMIP problem to construct a upper bound of the optimal solution to the original NLMIP problem (1.41). Specifically, let δ_0^* be the optimal solution to (1.41) where f is replaced by $f(\mathbf{0})$ (so it becomes a LMIP problem), clearly we have that $J(\tau(\delta^*)) \le J(\tau(\delta_0^*))$. We will see the quality of such a upper bound *in our numerical examples in Section 5. Also note that if $J(\tau(\delta_0^*)) < J(\tau(\mathbf{0}))$, then the sufficient condition for toll reduction stated in Theorem 1.3 must be true.*

Remark 1.5 *By reformulating the path-link-based problem (1.37) to node-link-based one (1.41), it is not hard to see that we can actually use the variables involved in (1.41) to design a polynomial time algorithm that characterizes either sets A_i $(i = 1, 2, 3)$ in Proposition 1.3. See the Appendix for details.*

Remark 1.6 *In problem (1.41) we only restrict $\delta \in \{0,1\}^{|W||M||A|}$. However, in reality there may be other constraints on δ. For example, the information provided to all the users from a certain region must be the same (i.e., $\delta_a^{w,m} = \delta_a \forall s_w \in N_1 \subset N,\ m \in M$), or only the area with higher accident rates or air pollution exposure should be considered to provide relevant safety or health risk information (i.e, $\delta_a^{w,m} = 0\ \forall a \notin A_1 \subset A,\ w \in W,\ m \in M$), etc.. Such constraints shrink the feasible region to $\delta \in \Delta \subset \{0,1\}^{|W||M||A|}$, and since these constraints are usually in trivial forms (such as those shown above), they allow a MTIDP with much smaller decision vector $\tilde{\delta}$ than δ by considering e.g., only region-based information selection, which makes the problem much easier to solve. These problem variants can be easily formulated in similar form as (1.41). We present here the most flexible information design problem with the largest solution space.*

Efficiently and exactly solving a NLMIP problem such as (1.41) is in general very difficult [63], meta heuristics are usually used in practice. In the next two subsections, we will present two practical algorithms to solve problem (1.41). The first one is a surrogate optimization (SUO) approach based on well-established method (e.g., [90]). The second one is a convex relaxation approach we specially designed for our MTIDP based on the problem structure.

1.4.2 A solution algorithm based on surrogate optimization

In many optimization problems, the performance of each feasible decision vector can be obtained by solving a mathematical programing problem or evaluating a "black box" simulation model. Hence sampling-based approach can be applied to solving this type of problems, which try to approach optimal solutions by evaluating a sequence of solutions one by one or group by group selected by some specially designed rules. Meta-heuristics (e.g., [35]) and Bayesian learning based algorithms (e.g., [45]) all belong to this type of "simulation optimization" approach. Solving our MITDP can also be viewed as a simulation optimization problem, since the performance of any binary decision vector δ, $J^*(\delta) = \min_{\tau \in \mathcal{T}(\delta)} J(\tau)$, can be evaluated by solving the dual problem (1.39) and the convex program $\min_{\tau \in \mathcal{T}(\delta)} J(\tau)$ (as more explicitly presented in Algorithm 1.1 below.

Algorithm 1.1: Evaluate $J^*(\delta)$ given δ

```
Solve the dual LP (1.39) under $\delta$ and obtain the optimal objective value OPT
Construct the linear inequalities characterizing $\mathcal{T}(\delta)$ (1.40) using the dual feasibility constraints in (1.39) and the restriction that the objective value of the dual LP (1.39) equals OPT
Solve problem (1.27) under $\delta$ and obtain an optimal solution $\tau^*(\delta)$ and optimal value $J(\tau^*(\delta))$
$J^*(\delta) \leftarrow J(\tau^*(\delta))$
```

Surrogate optimization (SUO) is a popular approach for solving simulation optimization problems. The idea of surrogate optimization is fitting an analytical model that approximates the complicated unknown relationship between system input (decision variables) and output (the performance measure) and searching for good solutions efficiently by the guidance of this surrogate model [90]. By the use of the surrogate model, the algorithm can usually accelerate the performance improvement along the sampling decisions and find the good solutions in relatively few evaluations. The SUO algorithms have been recently applied to transportation network optimization problems with satisfactory solutions found [27, 39].

For the MTIDP, we use the commonly used radial basis function (RBF) [101], we denote $\delta_1, ..., \delta_n \in \{0, 1\}^{|W||M||A|}$ the n sample points where the objective function $J^*(\delta)$ has been already evaluated, and let $\mathcal{D}_n = \{\delta_1, ..., \delta_n\}$. An RBF interpolant that is used to approximate the true function form can be expressed as (where $||\cdot||_2$ denotes the Euclidean norm)

$$r(\delta) = \sum_{i=1}^{n} \rho_i \phi(||\delta - \delta_i||_2) + b^{\mathrm{T}}\delta + c, \quad J^*(\delta) = r(\delta) + \epsilon(\delta), \tag{1.43}$$

where r represents the response surface (surrogate model); ϵ is the difference between the true system and the surrogate model output (model error); ϕ denotes the radial basis function (here we use the cubic RBF: $\phi(x) = x^3$); $b^{\mathrm{T}}\delta + c$ represents the affine tail term with $b = [b_1, ..., b_{|W||M||A|}]^{\mathrm{T}} \in \mathbb{R}^{|W||M||A|}$ that has the same dimension with the decision vector δ, and $c \in \mathbb{R}$. The parameters $\rho_1, ..., \rho_n$, $b_1, ..., b_k$ and c are determined by solving the following linear system of equations [58]

$$\begin{bmatrix} \Phi & D \\ D^{\mathrm{T}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \rho \\ \theta \end{bmatrix} = \begin{bmatrix} J \\ \mathbf{0} \end{bmatrix}, \text{ with } D = \begin{bmatrix} \delta_1^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ \delta_n^{\mathrm{T}} & 1 \end{bmatrix}, \tag{1.44}$$

where entry square matrix $\Phi \in \mathbb{R}^{n\times n}$ has its entry $\Phi_{ij} = \phi(||\delta_i - \delta_j||_2), i, j \in \{1, ..., n\}$, $\rho = [\rho_1, ..., \rho_n]^{\mathrm{T}}$, $\theta = [b_1, ..., b_{|W||M||A|}, c]^{\mathrm{T}}$, $J = [J^*(\delta_1), ..., J^*(\delta_n)]^{\mathrm{T}}$. The system (1.44) has a unique solution if and only if rank(D)=$|W||M||A| + 1$[101].

In general, the procedure of the surrogate optimization algorithms contain the following main steps: 1) Generate the initial experimental design; 2) Generate the set of candidate solutions 2) pick one of the candidate solution (that is not evaluated before) based on the surrogate model performance as the next sample solution; 3) Iterate until the total sampling budget N_{sample} is exhausted. In particular we implement the following SUO based algorithm 1.2 for MTIDP, which contains a subroutine (Algorithm 1.3) corresponds to step 2) and 3).

1.4.3 A solution algorithm based on semidefinite relaxation

Now we introduce a semidefinite relaxation (SDR) based approach to solve problem (1.41). SDR is by definition, a relaxation of the original nonconvex program to a semidefinite program (a type of general convex optimization problem). SDR is a powerful framework to deal with a family of nonconvex optimization problems, it has been successfully applied to many NP-complete combinatorial optimization problems [84] and many challenging problems in many areas such as signal processing and communications, [84]. Since a semidefinite program can be solved very efficiently by methods such as self-dual type of algorithms available in many solver packages [31], we just need a way to construct a good solution to the original NLMIP problem (1.41) based on the optimal solution of the SDR problem. In the rest of this subsection, we will briefly describe the idea of SDR and derive a SDR formulation for our problem (1.41).

First consider the following standard real-valued homogeneous quadratically constrained quadratic program (QCQP) in vector z:

$$\begin{aligned} \min_{z \in \mathbb{R}^n} \quad & z^{\mathrm{T}} C z \\ \text{s.t.} \quad & z^{\mathrm{T}} F_i z \leq f_i, \ i = 1, ..., l, \end{aligned} \tag{1.45}$$

Algorithm 1.2: SUO based algorithm for solving MTIDP (1.41)

```
Require: The input data of (1.41), $N_{sample}$, and $N_c$ (which is set to 500)
$D \leftarrow []$, $\mathcal{D} \leftarrow \emptyset$, $\delta_1 \leftarrow \mathbf{0}$,
Evaluate $J^*(\delta_1)$ by Algorithm 1.1, $J^{best} \leftarrow J^*(\delta_1)$, $\delta^{best} \leftarrow \delta_1$
while rank($D$) < $|W||M||A| + 1$ do
    Generate points $\{\delta_2, ..., \delta_{n_0}\}$ (where $n_0 = |W||M||A| + 1$) using the standard
    Latin hypercube design with in $[0, 1]^{|W||M||A|}$ and rounding to integers
    $D \leftarrow [\delta_1^{\mathrm{T}}, 1;\ ...\ ;\ \delta_n^{\mathrm{T}}, 1]$
end while
for $i = 2$ to $n_0$ do
    Evaluate $J^*(\delta_i)$ by Algorithm 1.1
    if $J^*(\delta_i) < J^{best}$ then
        $J^{best} \leftarrow J^*(\delta^*_{n_0})$, $\delta^{best} \leftarrow \delta^*_{n_0}$
    end if
end for
$J \leftarrow [J^*(\delta_k),\ k = 1, ..., n_0]^{\mathrm{T}}$
for $i = n_0 + 1$ to $N_{sample}$ do
    Compute $\rho$ and $b$ by solving (1.44) using $D$ and $J$, update $r(\cdot)$ in (1.43)
    Pick the next sample $\delta^*_i$ by Algorithm 1.3 using $\delta = \delta^{best}$, $\mathcal{D} = \mathcal{D}_{i-1}$ and $r(\cdot)$
    Evaluate $J^*(\delta^*_i)$ by Algorithm 1.1
    $D \leftarrow [D;\ \delta^*_i, 1]$, $J \leftarrow [J;\ J^*(\delta^*_i)]$, $\mathcal{D} = \mathcal{D} \cup \{\delta^*_i\}$
    if $J^*(\delta^*_i) < J^{best}$ then
        $J^{best} \leftarrow J^*(\delta^*_i)$, $\delta^{best} \leftarrow \delta^*_i$
    end if
end for
return $\delta^{best}$ and $J^{best}$
```

Algorithm 1.3: The sample decision step of SUO based Algorithm 1.2

```
Require: Current best point $\delta$, sampled points $\mathcal{D}$, and response surface $r(\cdot)$
Initialize $C^* \leftarrow \emptyset$
while $C^* \setminus \mathcal{D} = \emptyset$ do
    $P \leftarrow \min\left\{\max\left\{\frac{5}{|W||M||A|}, 0.1\right\}, 1\right\}$
    for $j = 1$ to $N_c$ do
        for $k = 1$ to $|W||M||A|$ do
            Pick $\zeta$ from 1,2,3 at random and draw $u \sim \mathcal{N}(0, \zeta^2)$
            $\bar{\delta}_k^j \leftarrow \delta_k^j + u$ and round to the nearest integer in $\{0, 1\}$
        end for
    end for
    Uniformly generate $N_c$ solutions $\bar{\delta}^{N_c+1}, ..., \bar{\delta}^{2N_c}$ in $\{0, 1\}^{|W||M||A|}$
    $C \leftarrow \{\bar{\delta}^1, ..., \bar{\delta}^{N_c}\} \cup \{\bar{\delta}^{N_c+1}, ..., \bar{\delta}^{2N_c}\}$
    $C^* \leftarrow \arg\min_{\delta \in C} r(\delta)$
end while
return $\delta^* \leftarrow$ any element in $C^* \setminus \mathcal{D}$
```

$$z^{\mathrm{T}} H_j z = h_j, \ \ j = 1, ..., m,$$

where l is the number of inequality constraints and m is the number of equality constraints. Matrices $C, F_1, ..., F_l, H_1, ..., H_m \in \mathbb{S}^n$ (where we denote the set of all real symmetric $n \times n$ matrices by $\mathbb{S}^n$), and $f_1, ..., f_l, h_1, ..., h_m \in \mathbb{R}$. Notice that the following important equalities hold by the basic properties of trace operation:

$$z^{\mathrm{T}} C z = \mathrm{Tr}(z^{\mathrm{T}} C z) = \mathrm{Tr}(C z z^{\mathrm{T}}),$$

$$z^{\mathrm{T}} F_i z = \mathrm{Tr}(z^{\mathrm{T}} F_i z) = \mathrm{Tr}(F_i z z^{\mathrm{T}}),$$

$$z^{\mathrm{T}} H_j z = \mathrm{Tr}(z^{\mathrm{T}} H_j z) = \mathrm{Tr}(H_j z z^{\mathrm{T}}),$$

which implies that the objective function and the constraints in (1.45) are linear in the matrix zz^{T}. Thus, by letting $Z = zz^{\mathrm{T}}$ and noting that $Z = zz^{\mathrm{T}}$ is equivalent to Z being a rank one symmetric positive semidefinite (PSD) matrix, we obtain the following equivalent formulation of problem (1.45):

$$\begin{aligned} \min_{Z \in \mathbb{S}^n} \quad & \mathrm{Tr}(CZ) \\ \text{s.t.} \quad & \mathrm{Tr}(F_i Z) \leq f_i, \ i = 1, ..., l, \\ & \mathrm{Tr}(H_i Z) = h_i, \ i = 1, ..., m, \\ & Z \succeq 0, \ \mathrm{rank}(Z) = 1, \end{aligned} \tag{1.46}$$

where $Z \succeq 0$ means Z is PSD. It is clear that the objective function and all the constraints in (1.46) are linear in matrix X, except for the constraint $\mathrm{rank}(Z) = 1$ (which is nonconvex). Hence the fundamental difficulty in solving (1.45) lies in this rank constraint. If we drop this rank constraint, we thus obtain a natural relaxation of problem (1.45):

$$\begin{aligned} \min_{Z \in \mathbb{S}^n, \ Z \succeq 0} \quad & \mathrm{Tr}(CZ) \\ \text{s.t.} \quad & \mathrm{Tr}(F_i Z) \leq f_i, \ i = 1, ..., l, \\ & \mathrm{Tr}(H_j Z) = h_j, \ j = 1, ..., m, \end{aligned} \tag{1.47}$$

which is a standard semidefinite program that can be solved efficiently.

We now show that our original NLMIP problem (1.41) can be written as a standard homogeneous QCQP in the form of (1.45), which enables us to formulate a SDR in the form of (1.47) for problem (1.41).

Proposition 1.4 *Define $l = |A|(1 + 2|W||M|)$, $m = n = 1 + |A| + (2|A| + |N|)|W||M|$, then there exit C, F_i, $H_j \in \mathbb{S}^n$ and f_i, $h_j \in \mathbb{R}$ $(i = 1, ..., l; \ j = 1, ..., m)$ such that problem (1.41) is equivalent to a QCQP in the form of (1.45).*

Proof: We construct vector $z = t[1, \hat{\delta}^{\mathrm{T}}, \tau^{\mathrm{T}}, f^{\mathrm{T}}, \eta^{\mathrm{T}}]^{\mathrm{T}}$, where $\tau \in \mathrm{R}^{|A|}$, $f \in \mathrm{R}^{|A||W||M|}$, and $\eta \in \mathrm{R}^{|N||W||M|}$ are defined earlier, and we define $t \in \mathbb{R}$ with $t^2 = 1$ and $\hat{\delta} \in \mathbb{R}^{|W||M||A|}$ with $\hat{\delta}_k = 2\delta_k - 1,\ k = 1, ..., |A|$. Hence $z \in R^n$. Note that for every $k = 1, ..., |W||M||A|$, $\delta_k \in \{0, 1\}$ if only if $\hat{\delta}_k^2 = 1$. Hence together with $t^2 = 1$ we first have $1 + |W||M||A|$ quadratic equality constraints in z. In problem (1.41) we have $|A|$ linear equality constraints in f^{T} that encode SO link flow condition, and another $|N||W||M|$ linear equality constraints in f^{T} which encode flow conservation condition. Finally, in problem (1.41) there are $|A||W||M|$ many equality constraints in $\hat{\delta}$, η, and τ that encode complementary slackness condition. All of these constraints can be translated into some quadratic constraints in z by noting that for any $A \in \mathbb{S}^n$ and $a \in \mathbb{R}^n$ (let $s = [\hat{\delta}^{\mathrm{T}}, \tau^{\mathrm{T}}, f^{\mathrm{T}}, \eta^{\mathrm{T}}]^{\mathrm{T}}$), it is true that

$$s^{\mathrm{T}} A s + a^{\mathrm{T}} s = 0 \ \Leftrightarrow \ z^{\mathrm{T}} \begin{bmatrix} 0 & a^{\mathrm{T}}/2 \\ a/2 & A \end{bmatrix} z = 0,\ t^2 = 1.$$

Therefore, we have a total of $m = 1 + 2|A| + (|A| + |N|)|W||M|$ quadratic equalities in z. Similarly, we can validate that the inequality constraints in problem (1.41) are equivalent to a total of $l = |A|(1 + 2|W||M|)$ quadratic inequalities in z, and the objective function in problem (1.41) is also quadratic in z. ■

Now suppose we obtain an optimal solution $Z_{opt} = zz^{\mathrm{T}}$ of problem (1.47) for the relaxed version of problem (1.41) , we want to recover a feasible and hopefully near optimal solution to the original problem (1.41). This is a crucial part of the SDR approach which is usually problem-dependent [84]. Note that we can approximate the PSD matrix Z^* by the following rank one matrix [84]

$$Z_{opt} = \sum_{i=1}^{n} \lambda_i q_i q_i^{\mathrm{T}} \approx \lambda_1 q_1 q_1^{\mathrm{T}} = (\sqrt{\lambda_1} q_1)(\sqrt{\lambda_1} q_1)^{\mathrm{T}}, \tag{1.48}$$

where $\lambda_i,\ i = 1, ..., n$, are the eigenvalues of matrix Z_{opt} sorted in an descending order, and $q_i \in \mathbb{R}^n,\ i = 1, ..., n$, are the corresponding eigenvectors. Thus if we first normalize the second to the $(1+|W||M||A|)^{\text{th}}$ entries in $\sqrt{\lambda_1}q_1$ by its last entry (corresponding to $t \in \{1,-1\}$) and then take the sign operation of these entires, we recover a vector $\hat{\delta} \in \{-1,1\}^{|W||M||A|}$. It follows that we can obtain a feasible solution z_1 to problem (1.41) by substituting $\delta = (\mathbf{1}+\hat{\delta})/2$ to the primal and dual problems (1.38) and (1.39) and solve these two problems.

We can further use randomization technique to improve the solution [84]. Specifically, in addition to z_1, we can randomly generate more feasible solutions $z_2, ..., z_{N_{sample}}$ of problem (1.41), where N_{sample} is some given number of samples. Let $\Sigma = Z_{opt}(1 : |W||M||A|+1, 1 : |W||M||A|+1)$ then the following relation is true:

$$\mathbb{E}_{\xi\sim\mathcal{N}(\mathbf{0},\Sigma)}\xi^{\mathrm{T}}\xi = \Sigma. \tag{1.49}$$

Hence a natural choice is to sample a vector $\xi \sim \mathcal{N}(\mathbf{0},\Sigma)$, then we divide the second to the $(1+|W||M||A|)^{\text{th}}$ entries in ξ by its last entry and take the sign operation of these entires, we recover a vector $\hat{\delta} \in \{-1,1\}^{|W||M||A|}$, finally we can solve for problem (1.27) with $\delta = (\mathbf{1}+\hat{\delta})/2$ and get a corresponding feasible solution to problem (1.41). We can repeat the random sampling $N_{sample}-1$ times, and together with z_1 we choose the one that yields the best objective J. This entire SDR based solution procedure is summarized in algorithm 1.4 below.

Theorem 1.4 *Both Algorithm 1.2 and Algorithm 1.4 converge to a global optimal solution δ^{opt} to problem (1.41) as N_{sample} increases.*

Proof: The convergence of the SUO based algorithm can be deduced by a simple counting argument [90]. We know that the number of feasible solutions δ are finitely many, and no solution will be sampled more than once, so the Algorithm 1.2 must sample the global minimum δ^* at some stage. The convergence of the SDR based algorithm lies in its randomization step. Specifically, because $\xi \sim$

$\mathcal{N}(\mathbf{0}, \Sigma)$ and the covariance matrix Σ has its diagonal entries $\Sigma_{ii} = 1$ as required by the constraints $\hat{\delta}_i^2 = 1$, and it is positive definite (as a result of interior point method in solving the SDP), so each $\hat{\delta}$ (and the corresponding binary vector δ after rescaling and taking the sign operation to each entry of $\hat{\delta}$) has positive probability to be chosen. Therefore, when $N_{sample} \to \infty$, the optimal solution δ^* will be selected and evaluated with probability 1. ■

The convergence result just shows finding an global optimal solution is ensured as N_{sample} increases. This is a weak result, as any brute-force search method can always find the optimal solution in finitely many steps since the solution space is finite and our problem is deterministic. However, the results of our numerical examples below show that the algorithms can approach good solutions (i.e., can find δ with $J^*(\delta)$ significantly less than $J^*(\mathbf{0})$) rather efficiently even when the solution space is very large.

1.5 Numerical Examples

To illustrate the results derived in the previous section and the performance of the solution algorithms we proposed, in this section we present two examples of MTIDP on two networks each with a different type of new information.

1.5.1 Example 1: Two-OD simple network

First, we consider a small benchmark network that was used by other studies to demonstrate the existence of SO-flow enabling tolls (e.g., [128, 130]). Here we use this network to convey the basic idea and modeling steps of the information

Algorithm 1.4: SDR based algorithm for solving MTIDP (1.41)

```
Require: The input data of (1.41) and $N_{sample}$
$l \leftarrow |A|(1+2|W||M|)$, $m, n \leftarrow 1+|A|+(2|A|+|N|)|W||M|$
Construct the coefficients $C,\ F_i,\ f_i, i=1,...,l$ and $H_j \in \mathbb{S}^n,\ h_j,\ j=1,...,m$ to the QCQP (1.45) formulated for (1.41) based on the problem data
Solve the relaxation of (1.45), (1.47), by interior point method to obtain an optimal solution $Z_{opt} \in \mathbb{S}^n_+$
$\lambda_1 \leftarrow$ largest eigenvalue of $Z_{opt}$, $q_1 \leftarrow$ eigenvector corresponds to $\lambda_1$
$q \leftarrow \sqrt{\lambda_1} q_1$, $t \leftarrow q(n)$, $\hat{\delta}_1 \leftarrow \mathrm{sgn}(q(2:|W||M||A|+1)/t)$
$\Sigma \leftarrow Z_{opt}(2:|W||M||A|+1, 2:|W||M||A|+1)$
for $i = 2$ to $N_{sample}$ do
    Sample $\xi_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$
    for $j = 1$ to $|W||M||A|$ do
        $\hat{\delta}_i(j) = \mathrm{sgn}(\xi_i(j)/t)$
    end for
end for
$J^{best} \leftarrow J^*(\mathbf{0})$, $\delta^{best} \leftarrow \mathbf{0}$
for $i = 1$ to $N_{sample}$ do
    $\delta_i \leftarrow (\mathbf{1} + \hat{\delta}_i)/2$
    Evaluate $J^*(\delta_i)$ by Algorithm 1.1
    if $J^*(\delta_i) < J^{best}$ then
        $J^{best} \leftarrow J^*(\delta_i)$, $\delta^{best} \leftarrow \delta_i$
    end if
end for
return $\delta^{best}$
```

design problem we introduced in section 1.3 as well as the effectiveness of the proposed toll intensity reduction algorithms. This network has four nodes and five directed links, as shown in Figure 1.2(a), where the link indices are highlighted beside the links. The network contains two OD pairs, one from A to D ($w = 1$), one from B to D ($w = 2$). Four paths are A $\rightarrow$ D ($p = 1$), A $\rightarrow$ C $\rightarrow$ D ($p = 2$), B $\rightarrow$ C $\rightarrow$ D ($p = 3$), B $\rightarrow$ D ($p = 4$). So $P_1 = \{1, 2\}$ and $P_2 = \{3, 4\}$. The link travel times (min) are given by affine functions $t_a(x_a)$ in the link flow variables x_a [130]: $t_1 = 20 + 2x_1,\ t_2 = x_2,\ t_3 = x_3,\ t_4 = 20 + x_4,\ t_5 = 2x_5$. In this example we assume the extra information on link a, e_a, is an estimate of the exposure to air pollutant NO_2 on the link, measured by the expected intake of NO_2 [21]

$$e_a(x) = \rho c_a(x_a)t_a(x_a),$$

where $\rho = 0.1$ is the average air pollution intake fraction by a commuter on the road, which is assumed the same for all the links. $c_a(x_a)$ is the NO_2 concentration on link a (ppb, parts per billion), as NO_2 is a relatively reactive gaseous species and sensitive to local traffic intensity, we assume c_a to be proportional to the traffic flow only on the link itself: $c_a(x_a) = 5(x_a),\ a = 1, ..., 5$. Suppose there are two user groups ($M = \{1, 2\}$) traveling on the network, their VOTs are: $\alpha_1 = 1,\ \alpha_2 = 2$, and we look at three choices of VOEs (valuation of exposure to NO_2): $\beta_1 = 0.1,\ \beta_2 = 0.2, \beta_1 = 0.2,\ \beta_2 = 0.1$ and $\beta_1 = \beta_2 = 0.15$.

We take the system optimal objective as the total delay on the network, i.e., $\Phi(x) = \sum_{a \in A} x_a t_a(x_a)$. Then we obtain the optimal link flow $x^* = [10,\ 10,\ 20,\ 10,\ 20]^{\mathrm{T}}$, which results in the link travel times $t_1^* = 40,\ t_2^* = 10,\ t_3^* = 20,\ t_4^* = 30,\ t_5^* = 40$ and the total travel time on the network is $\Phi(x^*) = 2000$. The expected average NO_2 exposure on the links under the optimal flow x^* are $e_1^* = 200,\ e_2^* = 50,\ e_3^* = 200,\ e_4^* = 150,\ e_5^* = 400$. When there is no provision of the air pollution exposure information ($\delta = \mathbf{0}$), the optimal solution to the primal LP

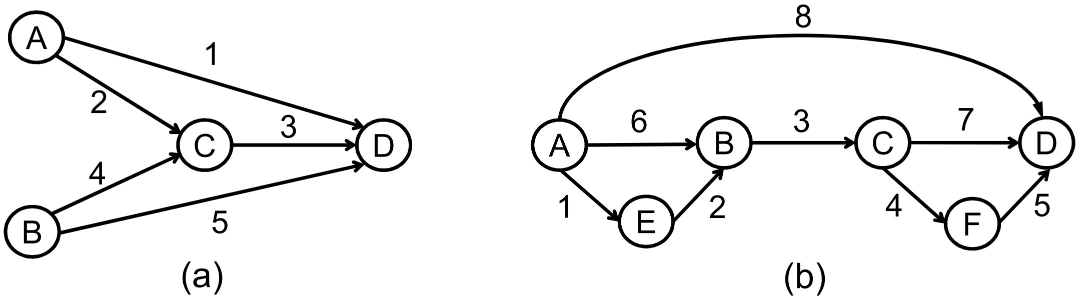


Figure 1.2: Two example networks (the numbers are link indices).

problem (1.21) is $f^* = [10,\ 0,\ 10,\ 10,\ 0,\ 10,\ 0,\ 10]^{\mathrm{T}}$, which means for OD pair 1, group 1 chooses path 1 while group 2 chooses path 2; and for OD pair 2, half users of group 1 chooses path 3 and the other half chooses path 4 while group 2 chooses path 4. The SO flow enabling tolls are contained in the polyhedron $\mathcal{T}(\mathbf{0})$ defined by the following linear constraints (derived from equation (1.25) with the variables γ_w^m eliminated):

$$\mathcal{T}(\mathbf{0}) = \left\{\tau \in \mathbb{R}_+^5 :\ 10 \le -\tau_1 + \tau_2 + \tau_3 \le 20;\ -\tau_3 - \tau_4 + \tau_5 = 10\right\}$$

The toll intensity function used is the total toll revenue, $J(\tau) = x^{*\mathrm{T}}\tau$. For each of the three VOE scenarios, Table 1.2 lists the info-selection vector in an optimal solution to problem (1.37), δ^{opt}, in the solution based on Theorem 1.3, δ^{rule}, as well as in the solution found based on one realization of the randomized SDR algorithm 1.4 (for the reformulation (1.41) of the original problem (1.37)) with sample size $N_{sample} = 5$, δ^{sdp}. Problem (1.47) is solved using the SeDuMi solver in CVX [31] (which is an implementation of interior point method). The corresponding optimal objective values of these solutions are also listed in Table 1.2. We denote $J^*(\bar{\delta})$ as the optimal value of problem (1.27) under $\delta = \bar{\delta}$, which is equal to the optimal value of (1.37) corresponding solution which has its info-selection vector $\delta = \bar{\delta}$.

Table 1.2: Optimal solutions for the first network example ($\alpha_1 = 1,\ \alpha_2 = 2$)

VOEs	**Scenario 1**	**Scenario 2**	**Scenario 3**
$[\beta_1, \beta_2]$	[0.1, 0.2]	[0.2, 0.1]	[0.15, 0.15]
Flexibility	**Only link selection allowed, i.e.,** $\delta_a^{w,m} = \delta_a\ \forall w, m$		
$J^*(\delta^{opt})$	150	0	75
Number of δ^{opt}'s	2	2	1
Selected links in δ^{opt}	(1,2,3,5); (1-5)	(3,4,5); (1-5)	(1-5)
$J^*(\delta^{rule})^{(a)}$	300	200	200
Selected links in δ^{rule}	None	2	2
Flexibility	**OD, link selection allowed, i.e.,** $\delta_a^{w,m} = \delta_a^w\ \forall m$		
$J^*(\delta^{opt})$	150	0	75
Number of δ^{opt}'s	64	48	32

Selected ODs and links in δ^{opt}, e.g.,	OD2:(3,5);
	OD2:(3,4,5)
		OD1:3,OD2(3,4,5);
		OD1:3,OD2:(2,3,4,5)
			OD1:2,OD2:(3,4,5);
			OD1:2,OD2:(2,3,4,5)

Flexibility	**User group, link selection allowed, i.e.,** $\delta_a^{w,m} = \delta_a^m\ \forall w$		
$J^*(\delta^{opt})$	0	0	0
Number of δ^{opt}'s	88	204	88

Selected groups and links in δ^{opt}, e.g.,	group2:(3,5);
	group2:(2,3,5)
		group2:(3,5);
		group2:(2,3,5)
			group2:(3,5);
			group2:(2,3,5)

Flexibility	**OD, user group, link selection all allowed**		
$J^*(\delta^{opt})$	0	0	0
Number of δ^{opt}'s	92160	194560	92160

Selected groups, ODs, and links in δ^{opt}, e.g.,	OD1-group2:(3),
	OD2-group2:(5);
	OD1-group2:(3),
	OD2-group2:(4,5)
		OD1-group2:(3),
		OD2-group2:(5);
		OD1-group2:(3),
		OD2-group2:(4,5)
			OD1-group2:(3),
			OD2-group2:(5);
			OD1-group2:(3),
			OD2-group2:(4,5)

[a] *The rules given in Proposition 1.3 only require link-based info. selection, so the same δ^{rule} and $J^*(\delta^{rule})$ also work for other flexibility cases.*

We can see that the toll intensity under no NO_2 exposure information is as high as 300 in order to minimize the total delay on the network. Provision of the air pollution exposure information can help lower this toll intensity. For example, disclosing the NO_2 intake information of all the links to all users under the second VOE scenario leads to zero toll intensity needed for minimizing the total delay. Under both the second and the third VOE scenarios with only link selection allowed in information design, the minimal toll vector that results from the provision only e_2 (based on the rule by Proposition 1.3) can cut the total toll charged by one third (specifically, link 2 is qualified for item (ii) in Corollary 1.1 under the second scenario and qualified for item (ii) in Corollary 1.1 under the third scenario). However, the rule in Proposition 1.3 does not help reduce the toll intensity for the first VOE scenario as in this case the sets $A_i,\ i = 1, 2, 3$ are all empty. Interestingly, under this network setting, the result of applying Corollary 1.1 is equivalent to applying Proposition 1.3. The results for other three information design flexibility cases verify that the reduction of SO-flow enabling toll intensity can be more significant when we have more flexibility in the information design. For example, if user groups can be selected in addition to the links in information scheme design, then we do not have to charge any user under the optimal solution δ^{opt}, i.e., $J^*(\delta^{opt}) = 0$, under all the three VOE scenarios. I.e., we can achieve optimal flow on the network purely by extra information provision. But note that adding one free dimension of OD in addition to link selection actually does not help improve the optimal values $J^*(\delta^{opt})$, which indicates the importance of having an optimization model that can help do the trade-off between policy flexibility choices and potential performance.

We also observe that although the number of feasible solutions increases exponentially in the size of the decision vector δ, including different free dimensions in information design lead to quite different proportions of optimal solutions out of the entire solution space. E.g., under the second scenario ($\beta_1 = 0.2$, $\beta_2 = 0.1$) the optimal solutions take up only $48/2^{10} \approx 5\%$ of all the feasible solutions with both OD and link selection allowed in the information design, which is lower than that ($2/2^5 = 6\%$) if only link selection is allowed. By contrast, under the same scenario, the optimal solutions take up $204/2^{10} \approx 20\%$ of all the feasible solutions with both user group and link selection are allowed, which is much higher than that ($2/2^5 = 6\%$) if only link selection are allowed. For the case with the most flexible information design (selection of OD, user group and links are all allowed) and also the largest solution space (2^{20}), we observe that there are about 9%, 19% and 9% solutions are optimal for scenario 1, 2, and 3, respectively. These proportions are almost the same with those when only user group and link selection considered and are significantly higher than other two flexibility cases, which indicates a potential of using solution algorithms that involve random sampling, such as the ones we proposed in the previous section.

Now we apply the two algorithms we introduced in the previous section to the MTIDP on this network. In order to test these algorithms, we choose the two flexibility cases where the proportion of optimal solutions among all the feasible solutions are the lowest according to the brute-force examination (see Table 1.2). These two cases are: 1) only link selection are allowed in information design; 2) both OD and link selections are allowed in information design. Figure 1.3 show the average performance of the two algorithms over 30 independent tests for each of the three scenario under various N_{smaple} values when only link selection is allowed in information design. Figure 1.4 show their performance when both OD and link selections are allowed. We can see from Figure 1.3 when only link selection is allowed, the value $J^*(\delta^{best})$ converges to the true global minimum rather fast using either algorithm for all three scenarios of different VOEs. In particular, we observe that on average, for any $N_{sample} \geq 2$ at least 25% toll intensity reduction (compared to $J^*(\mathbf{0}) = 300$) can be achieved by disclosing extra information on the set of links selected by the SDR based algorithm. By contrast, the SUO based algorithm needs $|A| + 1 = 6$ points for initial experiment design, but even when it stars with $N_{sample} = 6$ (i.e., only using random Latin hypercube sampling), the toll intensity reduction is quite notable. In addition, this reduction becomes more significant as the sample size N_{sample} increases. E.g., when N_{sample} reaches 10, on average the toll intensity reductions under the best solutions found by both algorithms are more than 90% of that achieved by the true optimal solution under all three scenarios of different VOEs.

When the flexibility of OD selection is also included in information design, we observe from Figure 1.4 that for the first two scenarios, the two algorithms can achieve about 90% of the maximum toll reduction from the baseline level $J^*(\mathbf{0}) = 300$ when N_{sample} reaches 25. But for the third scenario ($\beta_1 = \beta_2 = 0.15$), both algorithms need more than 40 samples to achieve good performance, this is intuitive as in this scenario only about $32/4^5 \approx 3\%$ feasible solutions are optimal, this percentage is the lowest among all three scenarios (see Table 1.2). We also observe that the SUO based algorithm seems to converge slightly faster than the SDR based approach for this scenario. We also noticed in our experiment (although not shown here) that under other two flexibility cases (user group and link selection allowed, and full flexibility case), where the proportions of optimal solutions are relatively higher, both algorithms can almost find an optimal solutions within 2 iterates (after the initial experimental design for the SUO base algorithm and during the random sampling for the SDR based algorithm).

Hence, the proposed solution methods are quite effective and efficient in founding high quality information schemes with various flexibility.

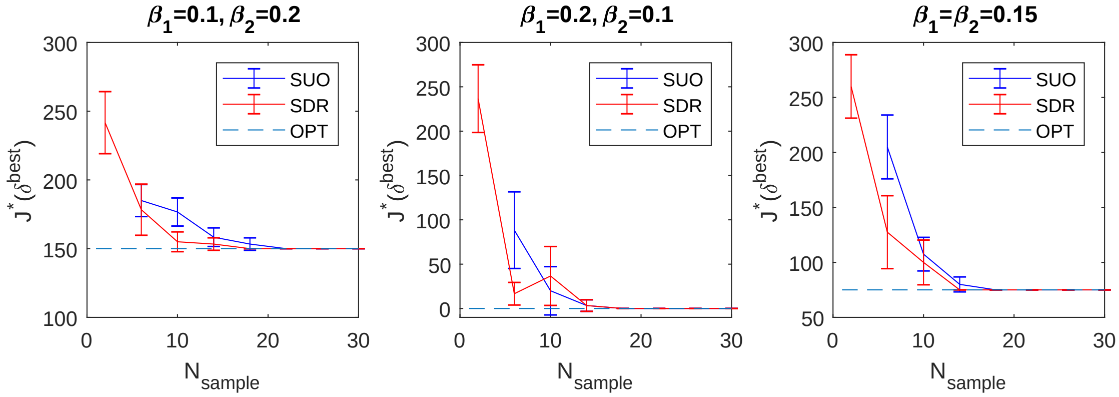


Figure 1.3: Average performance of Algorithm 1.2 (SUO) and Algorithm 1.4 (SDR) for the first network (only link selection allowed in information design). Each curve represents the sample mean of 30 independent tests, the error bar represents the 95% confidence interval for the mean estimate. The dash line represent the value of the true optimal solution $J^*(\delta^{opt})$.

1.5.2 Example 2: Series-parallel network

The second example is a series-parallel (SP) network that represents a common setting in traffic planning. This network has six nodes and eight directed links, as shown in Figure 1.2(b). We only consider one OD pair: node A to B, and two user groups $M = \{1, 2\}$ with demand $d_1^1 = d_1^2 = 10$. Note that link 1 to 7 can be regarded as local network available for those who choose to drive themselves. Link 3 represents a bottleneck link (e.g, a single bridge or tunnel) through which anybody who uses the local network must pass in order to start from node A to B. Link 8 represents a separate public transit route (e.g., subway line) from note

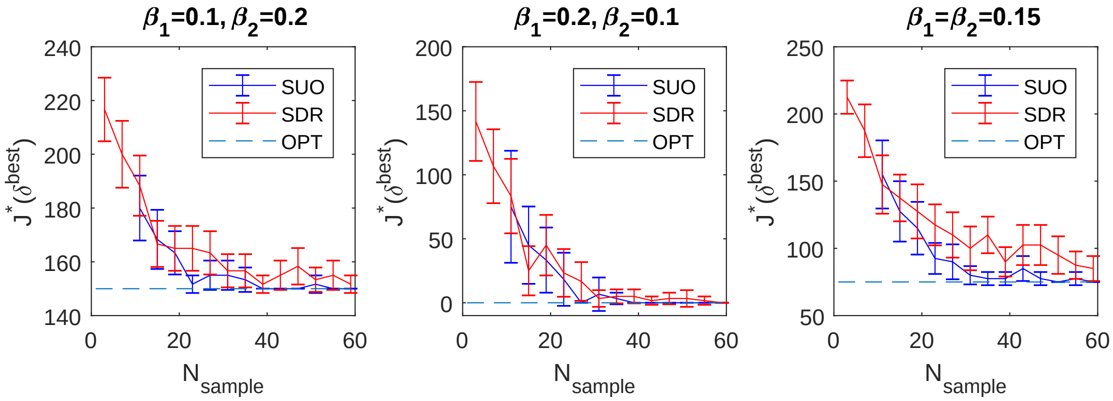


Figure 1.4: Average performance of Algorithm 1.2 (SUO) and Algorithm 1.4 (SDR) for the first network (both OD and link selection allowed in information design). Each curve represents the sample mean of 30 independent tests, the error bar represents the 95% confidence interval for the mean estimate. The dash line represent the value of the true optimal solution $J^*(\delta^{opt})$.

A to B. The link delay functions follows a classic BPR formula in the form [97]

$$t_a(x_a) = t_a^0 + \alpha_a \left(1 + \left(\frac{x_a}{c_a}\right)^{\beta_a}\right), \tag{1.50}$$

where t_a^0 is the free-flow travel time, $\alpha_a, \beta_a \geq 0$ are two parameters, c_a is the link capacity. Among the local network links, links 3, 6 and 7 have relatively larger capacity than links 1, 2, 4 and 5. These parameters for all the links are listed in Table 1.3. In this example, the extra information we focus on is the travel time variance on each link, which provides some indication of the travel time reliability. Since the separate transit line has no interaction with other traffic, we assume it has quite stable travel time with variance $e_8 = 0.25$. However, the local roads are subject to non-recurrent congestions caused by interactions of traffic, so in general the travel time uncertainty increases in travel time [85]. We assume that the travel time variance $e_a = 4t_a$ (in magnitude), $\forall a \in \{1, 2, 3, 4, 5, 6, 7\}$. Note that here $\beta_m > 0$ implies that users are all risk averse with utility contains a linear combination of mean plus variance of the travel time, this functional form was also adopted in other network routing and travel choice model studies (e.g., [88, 100]). In this example, we also use total delay as the SO objective, ϕ. The

Table 1.3: Link delay function parameters for the second network example

link a	1	2	3	4	5	6	7	8
t_a^0	0.1	0.1	0.2	0.1	0.1	0.2	0.2	1.25
c_a	5	5	10	5	5	10	10	20
α_a	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0
β_a	6	6	5	6	6	5	5	1

SO-flow and the original UE flow (i.e., $\delta = \mathbf{0},\ \tau = \mathbf{0}$) are

$$\begin{aligned} x^* &= [3.936, 3.936, 11.662, 3.916, 3.916, 7.726, 7.746, 8.338]^{\mathrm{T}}; \\ x^{UE} &= [5.481, 5.481, 16.633, 5.485, 5.485, 11.152, 11.148, 3.367]^{\mathrm{T}}. \end{aligned}$$

The total delay $\phi(x^{UE}) = 24.96$, which is 33.7% higher than $\phi(x^*) = 18.67$, because more users choose to drive themselves under the UE flow than that under the SO flow: under UE flow only 16.8% users choose transit while this number is 41.7% under the SO flow. The toll intensity function used for this example is the same as the first example, which is total delay. Thus, minimizing the total delay will require some users (at the UE flow) to switch from local routes that uses links 1~7 to link 8 (the transit line). Hence if we expect that when there is no extra information ($\delta = \mathbf{0}$), the toll should be applied to some of the links 1~7 to decentralize an SO flow. Compared to the first simpler example, this second example shows a trend that the number of paths can increase rapidly in the network size (here this number is 5). Hence we solve either (1.38) or (1.39) to characterize $\mathcal{T}$ according to (1.40). Then we solve problem (1.27) and get $\tau^*(\mathbf{0}) = [0.03, 0, 0, 0, 5.45, 0, 5.42, 0]^{\mathrm{T}}$ with $J(\tau^*(\mathbf{0}) = 63.42$, so there is a significant toll on the local roads to encourage self-drivers to use transit line instead, which is consistent to our expectation.

To demonstrate how travel time variability information can play a role in reducing the original toll intensity $J^*(\mathbf{0}) = 63.42$, we look at the optimal solutions to the MTIDP under four different scenarios of VOT and VOE distributions. In the first three scenarios, VOTs are the same: $\alpha_1 = 10$, $\alpha_2 = 20$. The first two scenarios seem plausible as the users who value expected travel time more also value travel time reliability more, and the ratio α_i/β_i are constant over all the users, which is 0.1 for scenario 1 and 0.2 for Scenario 2. In Scenario 1, if we only allow link selection in information design, the optimal solution is disclosing reliability measure on every local road to all the users, which cuts the tolls on local roads by more than 50%. This is intuitive, as users all hate the risk of being delayed due to longer travel time than expected, the provision of the travel time variation information on links 1~7 to users can help cut the toll needed for decentralizing a minimum-delay flow pattern if the Nash flow under $\tau^*(\mathbf{0})$ is unchanged when additional information is provided. Indeed, we verify that the Nash flow $f(\mathbf{0})$ (under $\delta = \mathbf{0}$ and $\tau^*(\mathbf{0})$ is also Nash under δ^{opt} and $\tau^*(\delta^{opt})$. Interestingly, if we also allow user group selection in information design, the corresponding optimal solutions show that giving the travel time reliability information of local roads only to users with lower VOT and VOE is enough, since actually 83.38% of this group of users choose to take the transit line under $f(\mathbf{0})$, which is exactly the SO flow on link 8. This means the new information and tolls only create incentive for group 1 to switch their route choices (taking transit instead of driving), which is enough to result in a SO flow. In Scenario 2, the VOE's are doubled compared to scenario 1, and this leads to a necessity in disclosing travel time variability on the transit line in addition to that on local roads (if all users need to receive the same set of information) in order to enforce a SO flow with least toll. The intuition behind this optimal solution is that as β_2 increases to 4 in this scenario compared to the previous one, group 2 also has an incentive to switch to transit line if they are only informed with the "risk" of extra delay on the local roads, which can lead to over-high flow on the transit line and worsen the efficiency. So the information e_a imposes an "balancing" effect to keep the most efficient flow, as we have verified that $f(\mathbf{0})$ remains a Nash flow under δ^{opt} and $\tau^*(\delta^{opt})$.

In Scenario 3, everything is the same with Scenario 1 except that $\beta_1 = -2$, which is negative, meaning users from group 1 become risk lovers [88]. This leads to a dramatic change in the optimal information schemes. As we checked that in this case $f(\mathbf{0})$ is still a UE flow under δ^{opt} and $\tau^*(\delta^{opt})$, and this means providing variability of travel time on transit lines, e_8, to user group 1 will encourage them to use link 8 instead of link $1 \sim 7$, and this is exactly reflected in the optimal solutions δ^{opt}, for both flexibility levels. Although seems unlikely, this example shows a caveat in optimal information design when some users behave very differently than other. If we had thought $\beta_1 = 1$ while in fact it is -2, we would have discouraged those risk lovers to switch to transit by not reporting the travel time variability on the transit line, we then need a much higher toll on the local road to make this up in order to maintain an SO flow. Therefore, our model provides an automatic way of capturing such possibilities and thus ensuring system efficiency by distributing information accordingly. By contrast, although the VOT and VOE values seem plausible and quite likely in the last scenario, the optimal information schemes δ^{opt} derived give us little intuition and are hard to imagine and reasoning. This illustrates that even a quantitative difference in system parameters can lead to very different and nontrivial optimal information distribution schemes. Therefore, an system optimization model such as the one we proposed is useful.

Table 1.4: Optimal solutions for the second network example

VOTs, VOEs	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**
$[\alpha_1, \alpha_2]$, $[\beta_1, \beta_2]$	[10, 20], [1,2]	[10, 20], [2, 4]	[10, 20], [-2, 2]	[12, 16], [3, 5]
Flexibility	**Only link selection allowed, i.e.,** $\delta_a^{w,m} = \delta_a \ \forall w, m$			
$J^*(\delta^{opt})$	30.49	3.38	57.59	0.27
Number of δ^{opt}'s	1	1	1	1
Selected links in δ^{opt}	(1-7)	(1-8)	(8)	(1,2,3,6,8)
Flexibility	**User group, link selection allowed, i.e.,** $\delta_a^{w,m} = \delta_a^m \ \forall w$			
$J^*(\delta^{opt})$	30.38	2.38	57.59	0.22
Number of δ^{opt}'s	2	4	4	6

Selected groups and links in δ^{opt}, e.g.,	group1:(1-7);
	group1:(1-7),
	group2:(8)
		group1:(1-7);
		group1:(1-7),
		group2:(8)
			group1:(8);
			group1:(8),
			group2:(8)
				group1:(3),
				group2:(1,2,3,5,6,8);
				group1:(3),
				group2:(1,2,3,4,6,8)

Figure 1.5 show the average performance of the two algorithms over 30 independent tests for each of the four scenario under various N_{smaple} values when both user and link selection are allowed in information design. We observed that our SDR based approach performs remarkably well under the first three scenarios: the Algorithm 1.4 can exactly find an optimal solution or nearly optimal solution starting from $N_{sample} = 1$. I.e., in these three scenarios, $\delta_1 = (\mathbf{1} - \hat{\delta}_1)/2$ is exactly an optimal info-selection vector (where $\hat{\delta}_1$ is extracted from $\lambda_1 q_1 q_1^{\mathrm{T}}$, the approximate rank one matrix of the SDP solution Z_{opt}) (see line 4 in Algorithm 1.4). The explanation of this is that these three scenarios all have $f(\mathbf{0})$ as a UE flow under δ^{opt} and $\tau^*(\delta^{opt})$, hence we expect that solving the problem is easier since in this case the NLMIP problem (1.41) can be reduced to a LMIP problem

(see Remark 1.4). By contrast, the SUO base algorithm performs just normally with the best objective found $J^*(\delta^{best})$ gradually approaching the optimal value, but with no sign of "early discovery". This interesting observation indicates that the proposed SDR base algorithm can sometimes uncover an optimal solution or near optimal solution very fast for certain problem instances due to the deep recognition of the problem structure of MTIDP (i.e., a QCQP), which can notably outperform surrogate model based sampling algorithms that only uses general response surfaces or meta-models fitted by the samples.

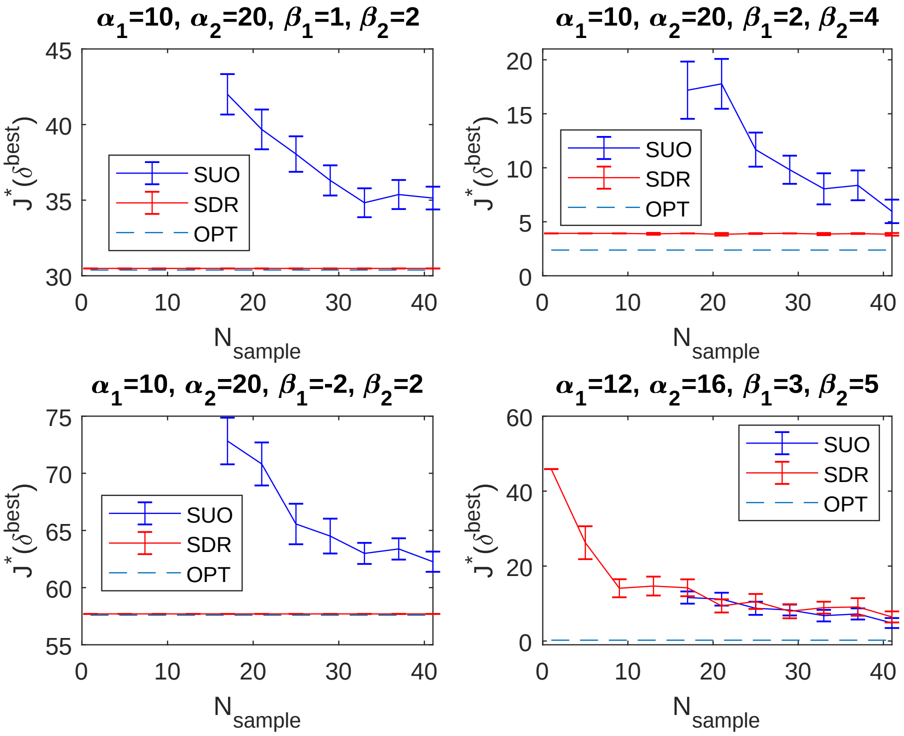


Figure 1.5: Average performance of Algorithm 1.2 (SUO) and Algorithm 1.4 (SDR) for the second network when both user and link selection are allowed in information design. Each curve represents the sample mean of 30 independent tests, the error bar represents the 95% confidence interval for the mean estimate. The dash line represent the value of the true optimal solution $J^*(\delta^{opt})$.

We test for robustness of our algorithm by solving 50 randomly generated instances of MTIDP using random VOT and VOE parameters for the network. Specifically, the VOTs are independent and uniform distributed as $\alpha_1 \sim \mathcal{U}(0,10)$, $\alpha_2 \sim \mathcal{U}(10,20)$, and VOEs are independent and uniformly distributed as $\beta_1 \sim \mathcal{U}(0,2)$, $\beta_2 \sim \mathcal{U}(2,4)$. I.e., we assume that on average each group of uses has notably smaller valuation to travel time reliability compared to expected travel time; and intuitively, users with higher VOT also have higher VOE. We test the case where both user and link selections are allowed in the information design. Figure 1.6 show the results. The upper two plots show the histograms of the ratio $J^*(\delta^{best})/J^*(\mathbf{0})$ obtained by two algorithms (averaged over 30 independent runs) over 50 independently generated random problem instances under two different choices of N_{sample}. The ratio $J^*(\delta^{best})/J^*(\mathbf{0})$ quantifies the relative toll intensity reduction potential under the best solutions found compared to the case with no reliability information), the smaller this ratio is, the more significant toll intensity reduction is expected. The lower two plots show the values of $J^*(\delta^{best})$ obtained by our algorithms (averaged over 30 independent runs) for each problem instance under two choices of N_{sample}. This values are put together with the true optimal value $J^*(\delta^{opt})$ and the initial minimal toll intensity needed $J^*(\mathbf{0})$ when there is no travel time reliability information.

We observe (from both the upper and lower plots) that $\sim$ 95% of the time, our two algorithms can find good information scheme solutions that result in decent toll intensity reduction. But in very few "bad" problem instances, the resultant information solutions actually increase the toll intensity needed for an SO flow. The average of the mean ratio $J^*(\delta^{best})/J^*(\mathbf{0})$ being 0.31 and 0.32 for the SUO based and SDR based algorithm, respectively, under $N_{sample} = 30$ and being 0.18 and 0.25 for the SUO based and SDR based algorithm, respectively, under $N_{sample} = 50$. Hence in average (over 50 problem instances) the SUO based algorithm achieves a slightly higher toll intensity reduction benefit than the SDR based approach. However, we note that each run of the the SDR based approach is much faster than the SUO approach, for example when $N_{sample} = 50$ ($N_{sample} = 30$), each run of the SUO Algorithm takes 9 ∼ 33s (3.5 ∼ 9s) while each run of the SDR Algorithm takes only 1 ∼ 6s (0.5 ∼ 6s), In addition, while the overhead of SUO approach is slowly increasing over the samples (due to increasing work in fitting the response surface), more than two thirds of the computational time for the SDR based Algorithm is spent on solving the SDP to first get a covariance matrix, each of the subsequent sampling is relatively efficient and independent of the total random sample size. Hence we expect that the SDP approach can have more improvement given the same computational time.

We want to point out that the ratio $J^*(\delta^{best})/J^*(\mathbf{0})$ can be used to bound the "sub-optimality"(in terms toll intensity reduction) of the best solution found by our algorithm, δ^{best}, by observing that

$$\frac{J^*(\mathbf{0}) - J^*(\delta^{best})}{J^*(\mathbf{0}) - J^*(\delta^{opt})} \geq \frac{J^*(\mathbf{0}) - J^*(\delta^{best})}{J^*(\mathbf{0})} = 1 - \frac{J^*(\delta^{best})}{J^*(\mathbf{0})}. \tag{1.51}$$

Therefore, the results of the above random instance test show that the performance of our algorithms are satisfactory.

Finally, via the same random instance test, we also compute the upper bound provided by an optimal solution δ_0^* to the reduced LMIP problem assuming $f(\mathbf{0})$ is still Nash (see Remark 1.4). Given the flexibility of both link and user selection in information design, Figure 1.7(a) shows the upper bound $J^*(\delta_0^*)$ compared with the true optimal value $J^*(\delta^{opt})$ and the original minimum toll intensity without information $J^*(\mathbf{0})$ over those 50 random problem instances, and Figure 1.7(b) plots the histograms of the ratios $J^*(\delta^{opt})/J^*(\mathbf{0})$ and $J^*(\delta_0^*)/J^*(\mathbf{0})$. It can be seen that the gap between the upper bound $J^*(\delta_0^*)$ and the true optimal value

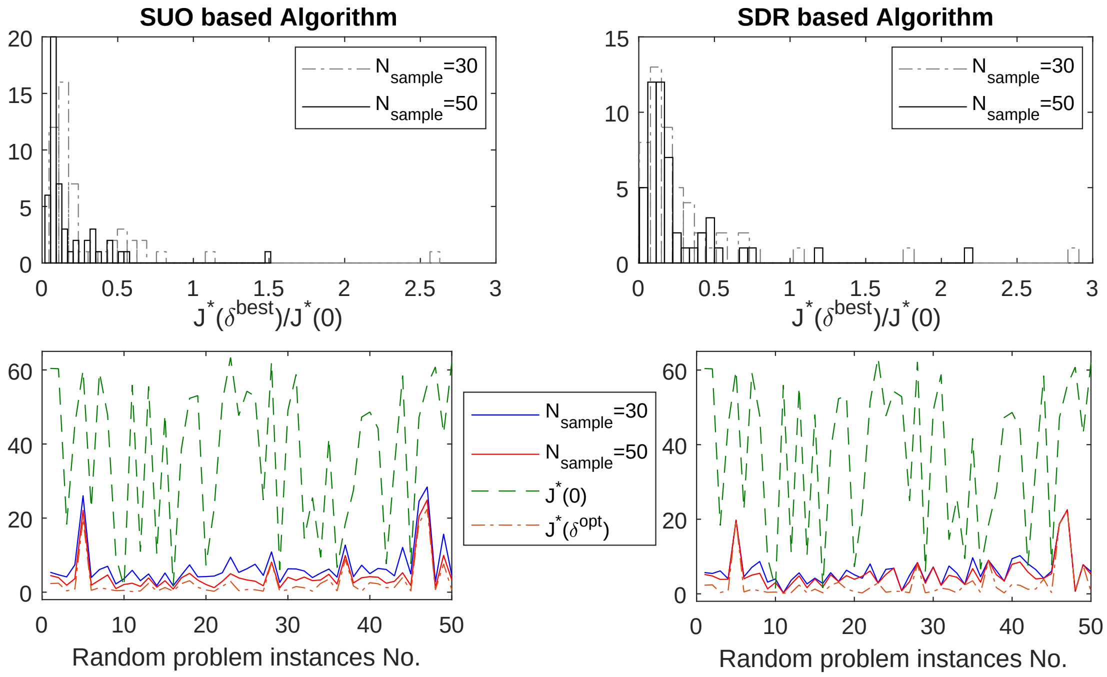


Figure 1.6: Performance of Algorithm 1.2 (SUO) and Algorithm 1.4 (SDR) for the second network over 50 random problem instances when both user and link selection are allowed in information design. Upper plots are the histograms of the average ratio $J^*(\delta^{best})/J^*(\mathbf{0})$ over 30 independent trials. Lower plots show three types of objective values for 50 problem instances: 1) $J^*(\delta^{best})$ (averaged over 30 runs, in red and blue lines); 2)$J^*(\delta^{opt})$; and 3) $J^*(\mathbf{0})$. (Each type of objective values for 50 problem instances are connected together).

$J^*(\delta^{opt})$ is notable for almost half of the problem instances, but for almost all the scenarios $J^*(\delta_0^*)$ is significantly lower (less than 60%) than $J^*(\mathbf{0})$. Therefore, in practice, even by solving the easier LMIP problem, we can expect a decent toll reduction by using the information design specified by δ_0^*.

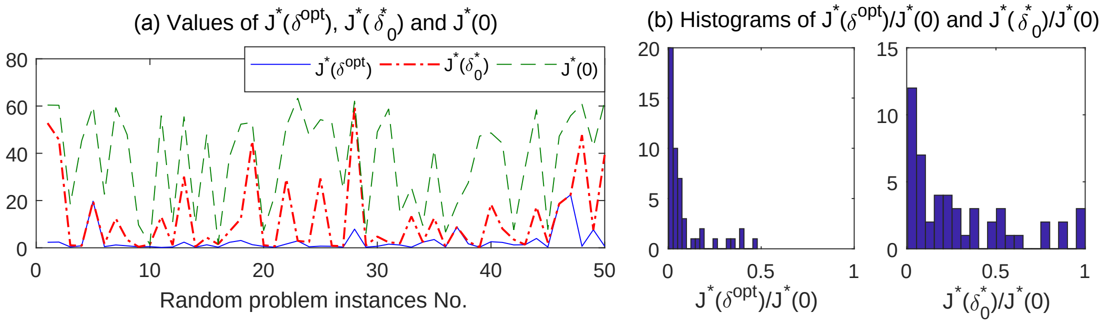


Figure 1.7: Comparison of the true optimal value $J^*(\delta^{opt})$, the upper bound $J^*(\delta_0^*)$ and the baseline value $J^*(\mathbf{0})$ for the second network over 50 random problem instances when both user and link selection are allowed in information design. (Each type of objective values for 50 problem instances are connected together in the left plot)

1.6 Conclusion and Extensions

In this study we have explored how strategic distribution of extra spatially resolved travel information can help reducing the toll intensity needed for achieving system-optimal flow on a traffic network. Based on the assumptions that user have an linear additive perceived cost function and the underlying quantity of the new information is link additive, We have proved a fundamental results characterizing the feasible toll set for enforcing an optimal flow under new information. We have discussed when "full information" may not be able to lower this toll intensity and when "partial information" is guaranteed to lead to positive toll intensity reduction. A general model of MTIDP was formulated together with two practical solution approach. Results of representative numerical examples show satisfactory performance of the proposed our algorithms in dealing with various restrictions in information design as well as distributions of user behavior parameters. Now we are experimenting on larger networks, for which each function evaluation by solving the large-sized LPs (Algorithm

1.1) can take up-to hours or even days. In this case, an efficient solution algorithm will be highly desired. We are also exploring how to scale up the SDR approach to efficiently obtain a initial covariance matrix Σ in the first place for larger applications.

There can be a number of natural and interesting extensions to this work. For example, instead of using total cost aggregated from link flows, we can define the SO objective in terms of user-weighted cost (e.g., total perceived dis-utility related to delay). We can add another dimension in the information design problem when more than one types of new information are available. In addition, as we can see from the numerical examples that the optimal information schemes are quite sensitive to the behavior parameters, so taking behavior and demand uncertainty into account in the design of robust information distribution schemes can be an interesting future exploration. Besides, model extensions can be studied to consider more realistic factors such as pre-system perceived cost and indirect information acquisition through subjective estimate of spacial correlation and network-effect among different groups of users.

1.7 Appendix

1.7.1 Proof of Proposition 1.3

Proof: The key is to prove that for one link $a \in A_0$ (i.e., $x_a^* > 0$) that is in either of the three sets A_i ($i = 1, 2, 3$), $\exists\, \tau^{new} \in \mathbb{R}^{|A|}$ with $J(\tau^{new}) < J(\tau)$ such that updating $\delta = 0$ by setting $\delta_a = 1$ (we denote δ^{new} the new info-selection vector) will keep $f(\mathbf{0})$ a UE flow. Then Theorem 1.3 implies that $\min_{\tau \in \mathcal{T}(\delta^{new})} J(\tau) < J(\tau(\mathbf{0}))$.

1) Suppose link $a \in A_1$, and we set $\delta_a^{new} = 1$ and $\delta_{a'}^{new} = 0\ \forall a' \neq a$. Then if we construct a toll vector τ^{new} such that $\tau_a^{new} = \max\{0, \tau_a - \beta_{\hat{m}_a} e_a^*\}$ and $\tau_{a'}^{new} = \tau(\mathbf{0})_{a'}\ \forall a' \neq a$, the perceived cost of any path p that contains link a by any group $m \in M$ does not decrease since

$$\begin{aligned}
\text{If } \beta_m e_a^* \geq \beta_{\hat{m}_a} e_a^* : \quad & g(\delta^{new})_p^m = g(\mathbf{0})_p^m + \beta_m e_a^* - \tau_a + \tau_a^{new} \\
& = g(\mathbf{0})_p^m + \beta_m e_a^* - \tau_a + \max\{0, \tau_a - \beta_{\hat{m}_a} e_a^*\} \\
& \geq g_p^m(\mathbf{0}) \geq \gamma(\mathbf{0})_w^m; \\
\text{Otherwise} : \quad & g(\delta^{new})_p^m = g(\mathbf{0})_p^m + \beta_m e_a^* - \tau_a + \tau_a^{new} \\
& \geq \gamma(\mathbf{0})_w^m + \min\{\tau_a, \beta_{\hat{m}_a} e_a^*\} - \tau_a + \max\{0, \tau_a - \beta_{\hat{m}_a} e_a^*\} \\
& = \gamma(\mathbf{0})_w^m,
\end{aligned}$$

and the perceived cost of any other path by any group is unchanged, i.e., $g_p^m(\delta^{new}) = g_p^m(\mathbf{0})\ \forall m \in M,\ w \in W,\ p \in P_w$ with $a \notin p$. Thus $f(\mathbf{0})$ (in particular, no users use link a) remains a UE flow under τ^{new} and δ^{new}.

2) Suppose link $a \in A_2$ is selected, $m_a \in M_a$ and we set $\delta_a^{new} = 1$ and $\delta_{a'}^{new} = \delta_{a'}\ \forall a' \neq a$. If we construct τ^{new} such that $\tau_a^{new} = \tau_a - \beta_{m_a} e_a^*$ and $\tau_{a'}^{new} = \tau_{a'}\ \forall a' \neq a$, then $f(\mathbf{0})$ remains Nash under δ^{new} and τ^{new}. Similarly, this is because the perceived cost on any path p that contains link a by any group $m \in M$ with $\beta_m e_a^* = \beta_{m_a} e_a^*$ does not change since by construction $\tau_a^{new} + \beta_{m_a} \delta_a^{new} e_a^* = \tau_a$, and for other groups $m \in M$ (i.e., with $\beta_m e_a^* \neq \beta_{m_a} e_a^*$),

$$\begin{aligned}
\text{If } \beta_m e_a^* > \beta_{m_a} e_a^* : \quad & g(\delta^{new})_p^m = g(\mathbf{0})_p^m + \beta_m e_a^* - \tau_a + \tau_a^{new} \\
& = g(\mathbf{0})_p^m + \beta_m e_a^* - \tau_a + \tau_a - \beta_{m_a} e_a^* \\
& > g(\mathbf{0})_p^m \geq \gamma(\mathbf{0})_w^m; \\
\text{If } \beta_m e_a^* < \beta_{m_a} e_a^* : \quad & g(\delta^{new})_p^m = g(\mathbf{0})_p^m + \beta_m e_a^* - \tau_a + \tau_a^{new} \\
& \geq \gamma(\mathbf{0})_w^m + (\beta_{\bar{m}} - \beta_m) e_a^* + \beta_m e_a^* - \tau_a + \tau_a - \beta_{m_a} e_a^*
\end{aligned}$$

$$= \gamma(\mathbf{0})^m_w,$$

and the perceived cost $g^m_p(\delta^{new}) = g^m_p(\mathbf{0})\ \forall m \in M,\ w \in W,\ p \in P_w$ with $a \notin p$.

3) Finally, suppose link $a \in A_3$ is selected and we set $\delta^{new}_a = 1$ and $\delta^{new}_{a'} = \delta_{a'}\ \forall a' \neq a$. If we construct τ^{new} such that $\tau^{new}_a = \tau_a - \beta_{\bar{m}_a} e^*_a$ and $\tau^{new}_{a'} = \tau_{a'}\ \forall a' \neq a$, then again, $f(\mathbf{0})$ remains Nash under δ^{new} and τ^{new}. This is because for any group $m \in M$ that has $\beta_m e^*_a = \beta_{\bar{m}_a} e^*_a$, the perceived cost on any path p that contains link a is not changed, since by construction, $\tau^{new}_a + \beta_{\bar{m}_a} \delta^{new}_a e^*_a = \tau_a$, and for the other groups $m \in M$ (i.e., with $\beta_m e^*_a \neq \beta_{m_a} e^*_a$),

$$\begin{array}{rl}
\text{If } m \in M_a & (\text{then } \beta_m e^*_a < \beta_{\bar{m}_a} e^*_a \text{ by definition of } \bar{m}_a): \\
& g(\delta^{new})^m_p = g(\mathbf{0})^m_p + \beta_m e^*_a - \tau_a + \tau^{new}_a \\
& = \gamma(\mathbf{0})^m_p + \beta_m e^*_a - \tau_a + \tau_a - \beta_{\bar{m}_a} e^*_a \\
& < \gamma(\mathbf{0})^m_p \leq g(\mathbf{0})^m_{p'}\ \forall p' \in P_w,\ w \in W^m_a \text{ with } a \notin p; \\
\text{If } m \notin M_a \text{ and } \beta_m e^*_a < \beta_{\bar{m}_a} e^*_a: & g(\delta^{new})^m_p = g^m_p(\mathbf{0}) + \beta_m e^*_a - \tau_a + \tau^{new}_a \\
& \geq \gamma(\mathbf{0})^m_w + (\beta_{\bar{m}_a} - \beta_m) e^*_a + \beta_m e^*_a - \tau_a + \tau_a - \beta_{\bar{m}_a} e^*_a \\
& = \gamma(\mathbf{0})^m_w, \\
\text{If } m \notin M_a \text{ and } \beta_m e^*_a > \beta_{\bar{m}_a} e^*_a: & g(\delta^{new})^m_p = g^m_p(\mathbf{0}) + \beta_m e^*_a - \tau_a + \tau^{new}_a \\
& = g(\mathbf{0})^m_p + \beta_m e^*_a - \tau_a + \tau_a - \beta_{\bar{m}_a} e^*_a \\
& > g(\mathbf{0})^m_p \geq \gamma(\mathbf{0})^m_w,
\end{array}$$

(in particular, the first of the above three cases indicates that under δ^{new} and τ^{new}, $\sum\limits_{p \in P_w : a \in p} f^m_p(\mathbf{0}) = d^m_w,\ \forall w \in W^m_a,\ m \in M_a$ with $\beta_m e^*_a \leq \beta_{\bar{m}_a} e^*_a$), and again $g^m_p(\delta^{new}) = g^m_p(\mathbf{0})\ \forall m \in M,\ w \in W,\ p \in P_w$ with $a \notin p$.

Note that in all of the above three scenarios (i.e., $a \in A_i$, $i = 1, 2, 3$), the constructed τ^{new} has $\tau^{new}_a < \tau(\mathbf{0})_a$ and $\tau^{new}_{a'} = \tau(\mathbf{0})_{a'}\ \forall a' \neq a$, thus by the property of toll intensity function (Assumption 1.5), it is true that $J(\tau^{new}) < J(\tau(\mathbf{0}))$. Therefore, the result follows by Theorem 1.3. ■

1.7.2 Proof of Corollary1.1

In this proof, the relevant variables and the sets A_i, $i = 1, 2, 3$ are defined in either the statement or the proof of Proposition 1.3.

First we proof item (i). Note that when $\beta_m = \beta \; \forall m \in M$ and $\beta e^*_a > 0$, A_1 and A_2 can be simplified to

$$A_1 = \{a : \tau(\mathbf{0})_a > 0,\ \mu_a = 0\},\ A_2 = \{a : \tau_a > 0,\ \mu_a = 1,\ \tau(\mathbf{0})_a \geq \beta e^*_a\}, \tag{1.52}$$

and $A_3 = \emptyset$ since $\mu_a \leq 1$. Thus a link $a \in A$ that qualifies conditions in (i) must lie in either A_1 or A_2. Now let's consider the argument given in Remark 1.2. Specifically, starting from $\delta = \mathbf{0}$ and $\tau = \tau(\mathbf{0})$, we scan over the links $a \in A$, each time we find a new link $a \in A$ that is qualified for (i) (i.e., in either A_1 or A_2 defined in (1.52), we update $\delta^{new}_a = 1$, $\delta^{new}_{a'} = \delta_{a'} \; \forall a' \neq a$, and we update the toll vector from τ to τ^{new} by reducing its entry corresponds to link a and keeping all the other entries unchanged (see the proof of Proposition 1.3). A key note here is that after such an update of δ_a (suppose $a \in A_i$), the set $A_i \setminus \{a\}$ is just the remaining links in the original set A_i defined in (1.52), because $\tau^{new}_a = \tau(\mathbf{0})_a \; \forall a \in A_i \setminus \{a\}$. As a result, since each update keeps $f(\mathbf{0})$ Nash and results in $J(\tau^{new}) < J(\tau)$. There fore, we have the result for case (i).

Then we show item (ii). We first consider two cases for a specific link a that is qualified for the conditions in (ii): 1) $\mu_a = 1$; 2) $\mu_a > 1$.

1) If $\mu_a = 1$, then we show that $a \in A_2$. Since a is used by all the shortest paths

of OD pairs $w \in W_a$, from Lemma 1.1 we know that the group $|M|$ (i.e., the group with the highest VOT) must uses link a under $f(\mathbf{0})$, and they must be the only group that uses link a under $f(\mathbf{0})$ since we know $\mu_a = 1$ and $\beta_m e_a^* \geq \beta_{m+1} e_a^* \ \forall m = 1, ..., |M| - 1$. Thus we deduce that $M_a = \{m \in M : \beta_m = \beta_{|M|}\}$ and $\beta_{m_a} e_a^* = \beta_{|M|} e_a^*$, it follows that the set A_2 reduces to

$$A_2 = \{a : x_a^* > 0, \ \tau(\mathbf{0})_a \geq \beta_{|M|} e_a^* > 0\}. \tag{1.53}$$

And by definition of m', we know $m' = |M|$, so $\tau(\mathbf{0})_a \geq \beta_{m'} e_a^* \leftrightarrow \tau(\mathbf{0})_a \geq \beta_{|M|} e_a^*$, so we deduce that $a \in A_2$.

2) If $\mu_a > 1$, then we show that $a \in A_3$. Firstly, since $\beta_m e_a^* \geq \beta_{m+1} e_a^* \ \forall m = 1, ..., |M| - 1$, by definition of $\bar{m}_a$ and m', we deduce that $m' = \bar{m}_a < |M|$. Then we deduce by Lemma 1.1 that $M_a = \{\bar{m}_a, \bar{m}_a + 1, ..., |M|\}$ and

$$\sum_{p \in P_w : a \in p} f_p^m = d_w^m \ \forall w \in W_a^m, \ m \in M_a \text{ with } \beta_m e_a^* < \beta_{\bar{m}_a} e_a^*.$$

since a is used by all the shortest paths between OD pair $w \in W_a$. Secondly, due to $M_a = \{\bar{m}_a, \bar{m}_a + 1, ..., |M|\}$ (as we derive above) and $\beta_m e_a^* \geq \beta_{m+1} e_a^* \ \forall m = 1, ..., |M| - 1$, we deduce that $\nexists \, m \in M \setminus M_a$ s.t. $\beta_m e_a^* < \beta_{\bar{m}_a} e_a^*$. Combining these two observations, we know the set A_3 is simplified to

$$A_3 = \{a : x_a^* > 0, \ \tau(\mathbf{0})_a \geq \beta_{\bar{m}_a} e_a^* > 0\}, \tag{1.54}$$

since $\bar{m}_a = m'$, so we deduce that $a \in A_3$.

Now consider looping over all of those links $a \in A$ that are qualified for (ii), we can verify that each update of δ to δ^{new} and τ to τ^{new} keeps $f(\mathbf{0})$ a Nash flow and results in $J(\tau^{new}) < J(\tau)$. And in addition, similar as case (i), after such an update of δ_a (suppose $a \in A_i$), the set $A_i \setminus \{a\}$ is just the remaining links in the original set A_i defined in (1.53) or (1.54), because $\tau_a^{new} = \tau(\mathbf{0})_a \ \forall a \in A_i \setminus \{a\}$. Hence we have the result. ■

1.7.3 Design of a polynomial time algorithm for Proposition 1.3

Here we discuss the key idea for designing a polynomial time algorithm that characterizes the sets A_i $(i = 1, 2, 3)$ defined in Proposition 1.3. Suppose $(f(\mathbf{0}),\ \tau(\mathbf{0}), \eta(\mathbf{0}))$ is an optimal primal and dual solution to problem (1.38) under $\delta = \mathbf{0}$. Here $f(\mathbf{0})_a^{w,m}$ is the flow of user group m between OD pair w on link a under $\delta = \mathbf{0}$; and $(\eta_n^{w,m} - \eta_{t_w}^{w,m})$ is the minimum generalized cost a user of group m would have incurred if she had started from node n to her destination t_w.

By the above definition of $f(\mathbf{0})_a^{w,m}$ $a \in A, m \in M, w \in W$, we know that

$$M_a = \{m \in M : f(\mathbf{0})_a^{w,m} > 0 \text{ for at least one } w \in W\}, \quad \sum_{p \in P_w : a \in p} f(\mathbf{0})_p^m = f(\mathbf{0})_a^{w,m}. \quad (1.55)$$

By the above definition of $\eta_n^{w,m}$ $n \in N, m \in M, w \in W$, we claim the following.

Lemma 1.3 *For some constant c, $g(\mathbf{0})_p^m - \gamma(\mathbf{0})_w^m \geq c \ \forall p$ with $a \in p$ is equivalent to*

$$\hat{c}(\mathbf{0})_{s_w,a_s}^m + \alpha_m t_a^* + \tau(\mathbf{0})_a + \eta(\mathbf{0})_{a_t}^{w,m} - \eta(\mathbf{0})_{s_w}^{w,m} \geq c, \quad (1.56)$$

where $\hat{c}(\mathbf{0})_{u,v}^m$ denotes the minimum generalized cost on any path from node u to v for user group m.

Proof: By definition of $\hat{c}(\mathbf{0})_{s_w,a_s}^m$, we know that $g(\mathbf{0})_p^m - \gamma(\mathbf{0})_w^m \geq c \ \forall p$ with $a \in p$ is equivalent to

$$\hat{c}(\mathbf{0})_{s_w,a_s}^m + \alpha_m t_a^* + \tau(\mathbf{0})_a + \hat{c}_{s_w,t_w}^m - \gamma(\mathbf{0})_w^m \geq c$$

since for any path $p \in P_w$ with $a \in p$, $g(\mathbf{0})_p^m \geq \hat{c}_{s_w,a_s}^m(\mathbf{0}) + \alpha_m t_a^* + \tau_a(\mathbf{0}) + \hat{c}_{s_w,t_w}^m$. We also know the minimum generalized cost $\gamma(\mathbf{0})_w^m = \eta(\mathbf{0})_{s_w}^{w,m} - \eta(\mathbf{0})_{t_w}^{w,m}$, then plug this into the LHS of the above relation, we have that

$$\hat{c}(\mathbf{0})_{s_w,a_s}^m + \alpha_m t_a^* + \tau(\mathbf{0})_a + \hat{c}(\mathbf{0})_{s_w,t_w}^m - \gamma(\mathbf{0})_w^m$$

$$
\begin{aligned}
&= \quad \hat{c}(\mathbf{0})^m_{s_w,a_s} + \alpha_m t^*_a + \tau(\mathbf{0})_a + \eta(\mathbf{0})^{w,m}_{a_t} - \eta(\mathbf{0})^{w,m}_{t_w} - (\eta(\mathbf{0})^{w,m}_{s_w} - \eta(\mathbf{0})^{w,m}_{t_w}) \\
&= \quad \hat{c}(\mathbf{0})^m_{s_w,a_s} + \alpha_m t^*_a + \tau(\mathbf{0})_a + \eta(\mathbf{0})^{w,m}_{a_t} - \eta(\mathbf{0})^{w,m}_{s_w},
\end{aligned}
$$

where the first equality is by definition of $\eta(\mathbf{0})^{w,m}_n$. Thus have the result. ■

Therefore, based on relation (1.55), Lemma 1.7.3, and any fast shortest path algorithm (for finding $\hat{c}(\mathbf{0})^m_{u,v}$ defined in (1.56)), we can clearly design a polynomial time algorithm to find sets $A_i,\ i = 1, 2, 3$.

CHAPTER 2

TRAFFIC STABILITY UNDER REAL-TIME EN-ROUTE AIR POLLUTION INFORMATION

In this chapter, we discuss the potential of providing specially chosen additional information in mitigating the negative effect on real-time routing caused by inaccuracy in travel time reporting. Our model analysis and experiment discussion augment the previous literature on traffic stability (mostly focus on the convergence to the user equilibrium flow under user choice adjustments) by adding the dimension of real-time flow dynamics under user choices and endogenous information feedback (with possibly varied accuracy).

Travel time information has been estimated and provided to drivers to help them make better routing decisions and alleviate congestion. However, because of challenges in data collection and sensor working principal, travel time information is often delayed and hence inaccurate. This inaccuracy can misguide motorists and result in unstable traffic patterns that exacerbate congestion. To alleviate this negative effect of travel time information on traffic flow, we explored the potential of providing drivers with real-time average en-route air pollution information (in addition to travel time). We developed a new queueing model that considers choice behavior of drivers provided with both travel time and air pollution information. Our model captures the impact of real-time air pollution information and the subsequent effects on traffic patterns. Results of our theoretical and numerical analysis indicate that provision of real-time air pollution information to travelers may help stabilize traffic. We further investigated how demand, choice behavior, emission and environmental parameters can affect this traffic stability enhancing effect. Such benefits are also demonstrated in our microscopic simulation of traffic on the George Washington Bridge using real-world data. We find that by posting real-time air pollution information, the average number of waiting vehicles and en-route vehicle stops can be reduced by as much as 69.8% and 17.1%, respectively, under a 5-min delay in travel time information. Our results shed light on a novel behavior-based transportation management strategy: informing drivers of real-time en-route air pollution information can assist dynamic routing, enhance traffic stability, and mitigate congestion. The proposed queueing model also provides a tractable way of studying real-time traffic dynamics under user choices and provision of endogenous information.

2.1 Introduction

2.1.1 Background and motivation

Traffic congestion and the associated air pollution is one of the most important challenges facing society today. In the United States alone, congestion in 2015 cost $160 billion in wasted fuel and time loss [108], which is about $1000 per commuter (and much more in metropolitan areas like Los Angeles and New York City) [125]. When the public health costs due to air pollution from vehicle emissions are included, such as asthma and respiratory illnesses, the total cost of traffic congestion is even much higher. The World Health Organization has estimated that around 3 million deaths each year worldwide are attributed to outdoor air pollution [127], to which motor vehicle emissions are a major contributor [136]. Pollution from fine particulate matter (referred to as $PM_{2.5}$), a key factor linked to premature deaths, is strongly associated with motor vehicle emissions. $PM_{2.5}$-related damage is estimated to be hundreds of billions of dollars per year worldwide [76].

To mitigate congestion and cope with growing demand for mobility, many cities have turned to “smart” traffic management by leveraging information technology. Adoption of various tools such as dynamic message signs (DMS) and smart-phone apps enable drivers to make better routing decisions. Moreover, the use of sensor networks allows for monitoring and improved use of the urban transportation infrastructure to an unprecedented extent. These tools are integrated into the Advanced Traveler Information System (ATIS), which provides travel time or/and traffic condition information to help drivers with their routing decisions [14, 74, 105]. The left plot in Figure 2.1 shows an example of a DMS near the west entrance to the George Washington Bridge in New York City, which displays the estimated travel times for the upper and lower levels of the bridge. Although travel time is the type of information most commonly provided by the ATIS, more general travel related information such as fuel/emission costs or air pollution levels can certainly be included [105].

A key issue in the provision of travel time information is that travel time estimates are usually lagged on account of the working principle of traffic detectors/sensors [34, 89] and/or post-processing/prediction algorithms [123]. Such lagged information does not well-reflect the real-time traffic condition and can misguide the drivers, which may create unstable flow distribution (according to the study of such service systems with heavy traffic [98]). In reality due to capacity constraints, unstable traffic distribution can worsen the congestion. The negative effect of inaccurate pre-trip travel information on user behavior and traffic distributions was observed and discussed by relevant studies (e.g., [14]), although the problem scale and focus are different from ours. In this study, we focus on real-time system performance and provide a model analysis that quantifies how much inaccuracy in the travel time reporting (measured by information delay here) the system can tolerate as a function of key parameters such as demand and user preferences. In addition, we propose providing real-time en-route air quality information as a remedy to the negative effect caused by delayed travel time information. This is based on the following considerations. First, sensor technologies have achieved significant advancements that make it practical to measure and disclose real-time air pollution information to travelers in a timely manner (e.g., 20 sec) [133, 139]; (by contrast, travel time estimates are usually delayed (e.g. by 15 mins) because of technical restrictions [34, 79]). Second, the concentration of certain pollutants such as ultra fine particles (UFP) [72] Thus in these senses, monitored on-road air pollution information can reflect more closely the real-time traffic condition. In addition, air pollution information can easily be included on DMS or in smart-phone apps and displayed to travelers (see the example in the right plot in Figure 2.1). More importantly, providing drivers with the air pollution information can help raise their environmental and health awareness, and thus has the potential to help them internalize the externalities of traffic congestion and emissions. Empirical evidence shows that people do value the availability of air pollution information that they can incorporate into their daily travel decisions (e.g., [8, 95]).

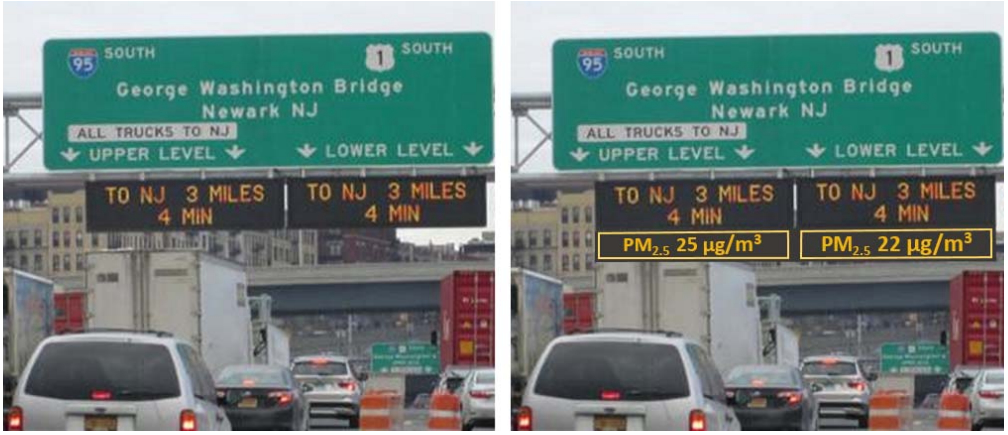


Figure 2.1: DMS at the George Washington Bridge (GWB). Travel time information (left panel), source: http://www.panynj.gov. The information of average $PM_{2.5}$ concentration on each level of the GWB are "added" artificially (right panel).

2.1.2 Literature review and our contribution

There has been a stream of literature on dynamic routing models under user choices, and in particular, the convergence to the equilibrium flow of these models (e.g., [23, 56, 115, 135]). For example, local and global stability of the equilibrium flow under day-to-day route choice adjustment process was analyzed in [135] based on previous experienced travel time. A link-based dynamical system model of day-to-day traffic adjustment on networks is studied in [56] with discussion of the properties of the system evolution including the stability of the equilibrium point. Authors in [23] studied the property of traffic equilibrium under node-by-node adaptive routing decisions based on new available real-time travel time information to the destination at each new node. Control measures such as dynamic pricing (e.g., [57]) or signal plans (e.g., [115]) in-response to day-to-day flow adjustment have also been proposed to facilitate the system convergence to the equilibrium flow on the traffic network. There are also a increasing number of studies on the effect of information provision on traffic flow distributions. For example, studies have shown via either theoretical analysis [80] or laboratory experiments [103] that under certain cases exogenous pre-trip information such as weather condition or road-work can worsen traffic distribution (in terms of social welfare) on parallel routes connecting a common origin-destination (O-D) pair. But they assumed that the information is accurate such that it is consistent with that experienced by the users. By contrast, authors in study [14] did an interesting experiment study on the impact of inaccurate pre-trip information on day-to-day route choice behavior, and useful implications for ATIS design were discussed.

However, there is no work to our best knowledge on how inaccurate information (such as delayed travel time estimate which is common in reality [34, 89]) can affect the real-time route choice dynamics and the resultant traffic flow patterns or proposed countermeasures to improve real-time system performance under inaccurate information. The stability of real-time adaptive node-to-node routing under travel time information was discussed in [23], but it was assumed that this information is accurate. In this study, we focus on discussing how delayed travel time information can cause undesired traffic patterns in routing choices. We show that given a set of competitive links serving fixed demand on the same O-D pair, delayed travel time information can cause unstable traffic, in which case the traffic distributions over different links keeps oscillating around the equilibrium. That is, the stale equilibrium flow (also referred to in the previous studies (e.g., [56, 135]) cannot be achieved. To remedy such negative effect, we also analyze the stability condition of the equilibrium flow under the provision of additional information that better reflects real-time traffic status.

The time delay problem is also found in many other service systems, such as call centers, hospitals, and amusement parks [98]. In these systems, queue length or waiting time information can be announced to customers for improved system management, but the announced information is usually estimated on the basis of past arrivals [65]. Thus, the development of relevant analytical methods to understand the impact of giving customers delayed information in such queueing services has attracted much attention from many research communities and is growing steadily (e.g., [65, 98, 99, 126]). A deterministic infinite-server queueing model with user choices was analyzed in recent studies [98, 99]. The authors showed that delays in information can have significant impacts on the performance of the system and provided insights into system stability under lagged queue length information. They found that the stability of the queue lengths can be characterized by a "critical delay" such that the queue lengths will converge to equilibrium when the information lag is smaller than this critical delay, but will oscillate indefinitely otherwise [98].

We generalize the queueing models of [98, 99] to consider a state-dependent service rate with more than one category of information provided to users. The state-dependent service rate is needed for representation of the relationship between traffic density and speed [68] and is usually explicitly included in dynamic routing models (e.g., [23]). We derive the critical delay under this new model and discover that traffic stability can be enhanced by providing travelers with real-time air pollution information in addition to an estimate of travel time. The air pollution information we adopted is in the form of the average en-route concentration of certain type of pollutant that is easy to measure and relatively sensitive to local traffic volume (as mentioned earlier).

In the rest of the paper we derive analytical results, conduct numerical and simulation experiments, and perform sensitivity analysis of key system parameters to illustrate the potential of disclosing real-time air pollution information in helping enhance traffic stability in real-time dynamic routing and mitigate congestion. To facilitate the derivation of the analytical results and description of the experimental data, we provide a notation table below that summarizes the main variables used in our model.

Variable	**Notation**	**Units**
Number of alternative links	N	/
Total vehicle arrival rate	λ	veh/min
Proportion of vehicles choosing link i	p_i	/
Number of vehicles on link i	Q_i	veh
Average travel time on link i	T_i	min
Delay in travel time information	Δ	min
Average air pollution concentration on link i	C_i	μg/m^3
Cross-sectional area of the "air box" for each link	A	m^2
Dilution rate constant of the pollutant	κ	min^{-1}
Background air pollution concentration	C_b	μg/m^3
Willingness to pay (WTP) of travel time saving	β_t	\$/min
WTP of air pollution concentration reduction	β_c	\$/(μg/m^3)
Travel time function in the number of vehicles Q_i	$g(Q_i)$	min
Emissions source strength function in Q_i	$h(Q_i)$	μg/m^3/s/veh
Total vehicle emission source strength on link i	S_i	μg/m^3/s

Table 2.1: Notation of the main variables in the model

2.2 Model

We consider a simple traffic network that has N links (indexed by $i \in \{1, \ldots, N\}$) serving traffic with fixed demand from a common origin region to a common destination region. At time t, both an estimate of the average travel time on each link with a common delay of Δ, $T_i(t-\Delta)$ (min), and average air pollutant concentration on each link of a certain air pollutant $C_i(t)$ (μg/m^3) are made available to drivers near the origin. Before presenting the details of our model, we made two main assumptions regarding the travel-time-related parameters and ambient environment conditions on alternative links, respectively.

Assumption 2.1 *The N alternative links are non-overlapping and physically separated with sufficient space in between. The N alternative links have the same length, fleet mix, free-flow travel speed and traffic capacity.*

Under this assumption, the alternative links have the same travel time function and emission source function in terms of the number of vehicles on the link (as will be defined more precisely later), and travelers cannot switch to the other link in the middle of the trip.

Assumption 2.2 *The N alternative links have the same emission dispersion condition and the same background air pollution concentration. The background concentration changes in a much longer time scale than the traffic dynamics and hence can be regarded as time invariant within the problem period to our interest (e.g., half a hour).*

Notice that the above two assumptions are made mainly for tractability of our theoretical model (the symmetry in the dynamical system equations) that we exploit in deriving the key insights. It well models the routing settings of competing alternatives such as parallel (but separate) tunnels or parallel (but separate) bridge links that connect to the same pair of origin and destination regions and have similar geometric configurations and ambient conditions. The symmetry condition does not necessarily hold in other real settings. However, our model can still be useful for understanding key qualitative implications in those settings. For instance, we will see that the phenomenon predicted by our analytical model is also observed in the simulation experiment where the symmetry assumption is relaxed (one link has larger traffic capacity than the other).

2.2.1 Traffic flow model

For purposes of traffic stability analysis, we study the traffic flow by use of queueing models. For a given total traffic demand represented by the total vehicle arrival rate λ (min^{-1}), we denote the arrival rate for link i by λ_i (min^{-1}). We assume that there are no connections between links (Assumption 2.1), so that once a motorist enters one link they cannot switch to another one.

To model the traffic flow on a link as a queueing system, we can think of the space occupied by an individual vehicle plus the headway between two adjacent vehicles as one "server, which goes into service at the time the vehicle enters the link and ends service when the vehicle reaches the end of the link [11]. We assume that the length L (m) of each link is much larger than the length of a server, so the number of servers on a link can be very large. We model the number of vehicles on link i as a state-dependent queueing process with a queue length of $Q_i(t)$ [11], so average vehicle density on link i is $\rho_i(t) = Q_i(t)/L$ (m^{-1}). A key feature of traffic flow is that an increase in vehicle density leads to a slowdown in the traffic, hence we assume that there is a positive, smooth, decreasing function $\mu(\cdot)$ such that the service rate on link i is $\mu_i(t) = \mu(Q_i(t))$ (min^{-1}) [68]. Function μ is the same for all links by Assumption 2.1. By this construction, the average travel time on link i is also a function $g(\cdot)$ of the queue length:

$$T_i(t) = \mu_i(t)^{-1} = \mu(Q_i(t))^{-1} := g(Q_i(t)). \tag{2.1}$$

2.2.2 Air pollution model

In addition the travel time, information on the measured air pollution can be posted to influence the routing choices of drivers. We denote the average pollutant concentration on link i by $C_i(t)$. As studies show, on-road or near-road air pollution is strongly affected by traffic. Since we assume that the links are spatially separated (Assumption 2.1), we model $C_i(t)$ explicitly in terms of the traffic density $\rho_i(t)$ or the number of vehicles $Q_i(t)$ on link i. Following a common approach used in the modeling of transient air quality [67], we construct an "air box" around the link segment that represents the space where the average concentration is measured (see Figure 2.2). Under given meteorological conditions, the average concentration in the air box evolves over time and depends on the emission source strength and the dispersion intensity, hence it can be modeled by the mass balance equation [29, 67]:

$$\dot{C}_i(t) = S_i(t) - \kappa(C_i(t) - C_b), \tag{2.2}$$

where $\kappa > 0$ is the pollutant dilution rate constant that represents the dispersion intensity (min^{-1}) [67]; κ depends mainly on meteorological conditions and the chemical/physical properties of the pollutant, so we assume it to be exogenous [29]. By Assumption 2.2, we use a common κ for the links. C_b is the background concentration (μg/m^3), which is assumed constant and common to all the links during the analysis period by Assumption 2.2. $S_i(t)$ is the emission source strength (μg/m^3/min) on link i, which is modeled as the product of the vehicle density $\rho_i(t)$, the average vehicle emission factor $r_i(t)$ (μg/m), the average vehicle speed $v_i(t) = L/T_i(t)$ (m/min), and the reciprocal of the cross-sectional area A (m^2) of the air box. Note that given the common meteorological conditions and fleet mix on the links, $r_i(t)$ can be approximated as a smooth function r of average speed $v_i(t)$ [136]: $r_i(t) = r(v_i(t))$. Thus by equation (2.1), the emission source strength $S_i(t)$ can be expressed as

$$\begin{aligned} S_i(t) &= \frac{r(v_i(t))v_i(t)Q_i(t)}{AL} = \frac{r(L/g(Q_i(t)))}{Ag(Q_i(t))} \cdot Q_i(t) \\ &= h(Q_i(t)) \cdot Q_i(t), \end{aligned} \tag{2.3}$$

where the function $h(x) = \frac{xr(L/g(x))}{Ag(x)}$ represents the average emission strength (μg/m^3/min) per vehicle in the queue, which is nonlinear. Specifically, for the functional form of $h(\cdot)$, we can substitute into equation (2.3) the empirically fitted emission factor function $r(\cdot)$ (in terms of speed) from the emission model and the travel time function $g(\cdot)$ (in terms of queue length) from the traffic model. Or we can directly fit a function $h(\cdot)$ using the queue length and data on the average vehicle emission strength by running the traffic and emission models together.

Therefore, by equation (2.3) we can re-write equation (2.2) as

$$\dot{C}_i(t) = h(Q_i(t)) \cdot Q_i(t) - \kappa(C_i(t) - C_b). \tag{2.4}$$

When the pollutant of concern is inert at the local scale (e.g., PM$_{2.5}$ or carbon monoxide (CO)) and the wind speed is low with stable atmospheric conditions, the value of κ is small, and thus C_i depends more on earlier emissions. When the pollutant is reactive (e.g. nitrogen oxides (NO$_\mathrm{x}$)) or wind speed is high with unstable atmospheric conditions, the emitted pollutants have higher dispersion rates [29, 67]. This latter case corresponds to higher values of κ, and C_i is less affected by earlier emissions and more representative of current emissions and traffic conditions.

Remark 2.1 *Equation (2.4) is defined for any type of pollutant emitted by vehicles to our interest. In our numerical and simulation experiments we use PM$_{2.5}$ as an illustrative example just because the data is more available for this type of pollutant. We want* *to emphasize that the same analysis and insights apply to other pollutants such as CO, NO_x and ultra fine particles (UFP). In fact, on/near road $PM_{2.5}$ level usually contains a large part from background concentration C_b that is comparable or higher than that contributed by the local traffic. But for those pollutants that are more dominated by local vehicle emissions or more reactive (such as UFP or NO_x), it is typically the case that emission source S is larger relative to C_b or κ is larger, so reporting the concentration level for such pollutants may have a bigger effect on routing choices and traffic distributions than $PM_{2.5}$ (as will be explained in detail in our discussion section). Thus the qualitative observations from our experiments also hold for other types of pollutants.*

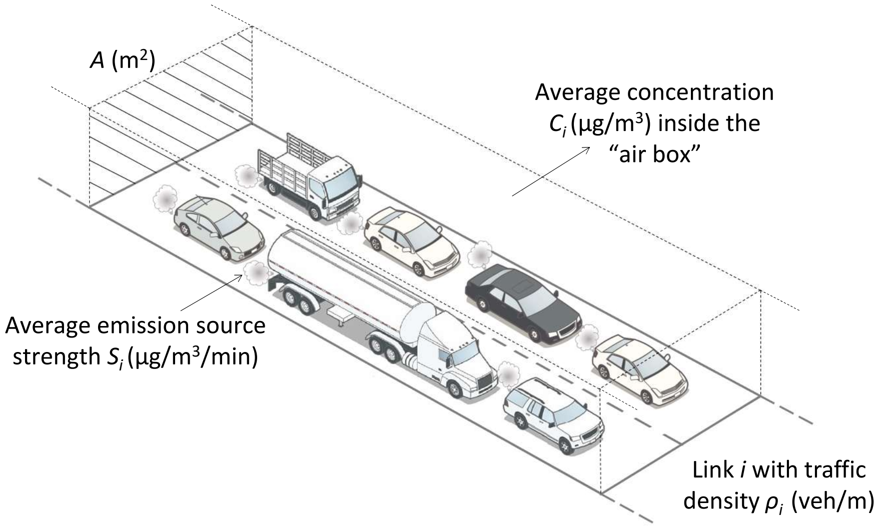

A (m^2)
Average concentration C_i (µg/m^3) inside the "air box"
Average emission source strength S_i (µg/m^3/min)
Link i with traffic density ρ_i (veh/m)

Figure 2.2: Vehicle queue and air concentration model for a link segment.

2.2.3 Route choice model

We use the multinomial logit (MNL) model to describe motorists' route choice dynamics. The MNL model is commonly used in the modeling of discrete choices in transportation [121]. In our MNL model, the perceived utility from information of average travel time with a delay of Δ, $T_i(t-\Delta)$, and the air pollutant concentration $C_i(t)$ measured real-time is $(-\beta_t T_i(t-\Delta) - \beta_c C_i(t))$, where $\beta_t > 0$ and $\beta_c \geq 0$ are motorists' marginal dis-utilities from the travel time and the air pollution, respectively. According to the economics literature, β_t and β_c are usually measured by the willingness to pay (WTP) [121] for the per-unit saving in travel time (\$/min) and per-unit reduction in the air pollution level (\$/(μg/m^3)), respectively. Hence we refer to β_t and β_c as WTPs in the sequel. Thus, the probability that a motorist will choose link i is

$$p_i(\mathbf{T}(t-\Delta), \mathbf{C}(t)) = \frac{\exp(-\beta_t T_i(t-\Delta) - \beta_c C_i(t))}{\sum_{j=1}^{N} \exp(-\beta_t T_j(t-\Delta) - \beta_c C_j(t))}, \tag{2.5}$$

where we define the vectors $\mathbf{T}(t-\Delta) = \{T_i(t-\Delta),\ i = 1, ..., N\}$ and $\mathbf{C}(t) = \{C_i(t),\ i = 1, ..., N\}$, respectively, as the delayed travel time information on the links and real-time air pollution concentration information on the links.

The main objective of our analysis was to study the traffic dynamics when both values of $\mathbf{C}(t)$ and $\mathbf{T}(t-\Delta)$ are provided. To do this, we analyzed a fluid model of traffic instead of the actual stochastic process, which is more difficult [98]. More importantly, the fluid model enabled us to study the mean dynamics of the traffic system when the number of arrivals is large [98, 99], which is usually the case for road traffic. Since the fluid model is deterministic, the choice probability given in equation (2.5) is the proportion of drivers that join link i. Thus, in our fluid queueing model, at time t the motorists join link i at the rate of (also a function of $\mathbf{T}(t-\Delta)$ and $\mathbf{C}(t)$)

$$\lambda_i(\mathbf{T}(t-\Delta), \mathbf{C}(t)) = \lambda p_i(\mathbf{T}(t-\Delta), \mathbf{C}(t)) = \frac{\lambda \exp(-\beta_t T_i(t-\Delta) - \beta_c C_i(t))}{\sum_{j=1}^{N} \exp(-\beta_t T_j(t-\Delta) - \beta_c C_j(t))}. \tag{2.6}$$

Using the models discussed above, we can analyze the stability of the traffic densities $\rho_i,\ i = 1, ..., N$, on the individual links. Since the traffic density on link

i is defined as $\rho_i = Q_i/L$, the results are similar to those that can be derived by analyzing the stability of the queue lengths $Q_i,\ i = 1, ..., N$. Thus we describe the results in queue lengths in the following section.

2.3 Results

In this section, we analyze our fluid queueing model to examine traffic stability. Of particular interest is the derivation of the critical delay in the travel time information as a function of the parameters β_t, β_c. If the lag in the posted travel time is less than the critical delay, the vehicle densities will gradually synchronize across all links and reach a stable balance; otherwise, the vehicle densities on different links remain asynchronous, i.e, the system is unstable.

2.3.1 Analytical results

By results from the basic queueing fluid model of traffic [98], together with the air pollutant concentration from equation (2.4) and the vehicle arrival rate from equation (2.6), we have the following system of symmetric $2N$-dimensional delay differential equations ($i = 1, ..., N$):

$$\begin{aligned}
\dot{Q}_i(t) &= \lambda_i(\mathbf{T}(t-\Delta), \mathbf{C}(t)) - \mu_i(t)Q_i(t) \\
&= \lambda p_i(\mathbf{T}(t-\Delta), \mathbf{C}(t)) - \mu_i(t)Q_i(t) \\
&= \frac{\lambda \exp\left[-\beta_t T_i(t-\Delta) - \beta_c C_i(t)\right]}{\sum_{j=1}^{N} \exp\left[-\beta_t T_j(t-\Delta) - \beta_c C_j(t)\right]} - \frac{Q_i(t)}{T_i(t)} \\
&= \frac{\lambda \exp\left[-\beta_t g(Q_i(t-\Delta)) - \beta_c C_i(t)\right]}{\sum_{j=1}^{N} \exp\left[-\beta_t g(Q_j(t-\Delta)) - \beta_c C_j(t)\right]} - \frac{Q_i(t)}{g(Q_i(t))},\ i = 1, ..., N; \\
\dot{C}_i(t) &= S_i(t) - \kappa[C_i(t) - C_b]
\end{aligned}$$

$$
\begin{aligned}
&= S(Q_i(t)) - \kappa[C_i(t) - C_b] \\
&= h(Q_i(t)) \cdot Q_i(t) - \kappa[C_i(t) - C_b],\ i = 1, ..., N, \qquad (2.7)
\end{aligned}
$$

where $T_i(t) = g(Q_i(t))$ and $h(\cdot)$ are the positive, smooth functions defined earlier.

Suppose there exists a positive root Q^* for function $q(Q_i) := \frac{g(Q_i)}{Q_i} - \frac{\lambda}{N}$ with $q'(Q^*) > 0$ (q' denotes the derivative of q), then the equilibrium solution of the system of equations (2.7) is

$$
\begin{aligned}
&Q_1 = Q_2 = ... = Q_N = Q^*; \\
&C_1 = C_2 = ... = C_N = \frac{1}{\kappa} h(Q^*) \cdot Q^* + C_b. \qquad (2.8)
\end{aligned}
$$

Now we will discuss the system of equations given in (2.7) using a numerical example where we have two parallel single-lane links ($N = 2$) of the same length, $L = 1$ (mile), and a free-flow speed of 60 (mile/hour). The fitted travel time in terms of the queue length is $g(Q_i) = 0.0526Q_i + 1.0$ (min). The average emission strength for $PM_{2.5}$ in terms of the queue length is $h(Q_i) = 1.176 \exp(-0.0058Q_i)$ (μg/m^3/min) for an on-road air box of cross-sectional area $A = 15$ (m^2). We used $C_b = 35$ (μg/m^3) for the background concentration, and $\kappa = 6$ (min^{-1}) for the dilution rate. The WTP for travel time was taken to be $\beta_t = 0.6$ (\$/min). Further details on these estimates are given in Appendix B. We assumed that the total arrival rate was $\lambda = 30$ (veh/min) and that the WTP for air pollution was $\beta_c = 0.2$ (\$/(μg/m^3)). The initial conditions we used are $Q_1 = Q^* + 20,\ Q_2 = Q^* - 20;\ C_i = h(Q_i) \cdot Q_i + C_b,\ i = 1, 2$. Figure 2.3 illustrates the stability outcome under two different time-lag scenarios for posting of travel time information: $\Delta = 4$ (min) and 6 (min). We can see that when the time lag is $\Delta = 4$ (min) the average traffic densities $\rho_1,\ \rho_2$ and the average $PM_{2.5}$ concentrations $C_1,\ C_2$ on the two links balance out after about $t = 100$ (min). However if the time lag Δ increases to 6 (min), the traffic densities and $PM_{2.5}$ concentrations continue to oscillate and don't converge to an equilibrium. It turns out in this example that the traffic densities and $PM_{2.5}$ concentrations will be synchronous if the time lag Δ is less than 5.425 (min); otherwise, the system will continue to oscillate (and never settle down). More generally, there exists a critical delay Δ_{cr} in the posting of travel time information such that the traffic densities on both links will converge to the equilibrium solution (2.8) if $\Delta < \Delta_{cr}$, but will keep oscillating otherwise.

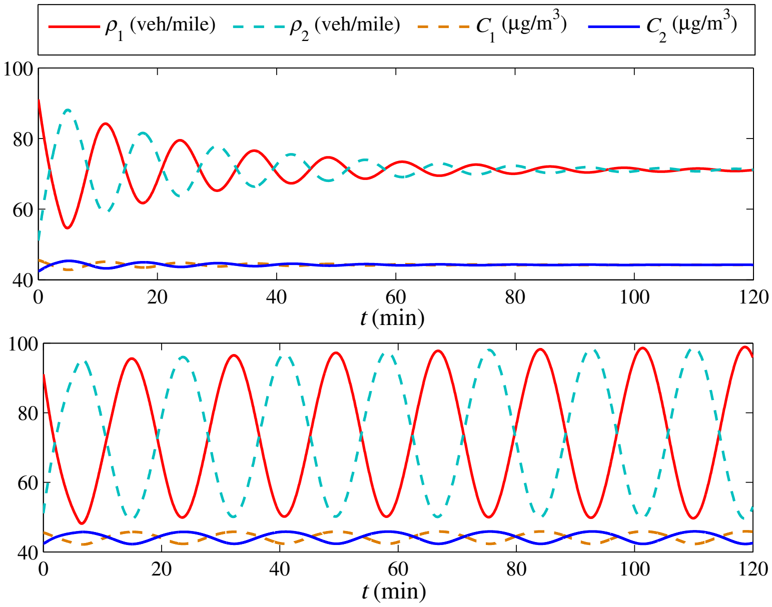


Figure 2.3: Example: $\lambda = 30$, $\beta_c = 0.2$, $\Delta = 4$ (upper), $\Delta = 6$ (lower).

To characterize the traffic stability we observed, we analyzed the system of nonlinear delay differential equations given in (2.7) via nonlinear dynamics. This leads to our first theorem below, the proof of which is provided in the Appendix A.

Theorem 2.1 *Define the following quantities:*

$$\xi \quad = \quad g'(Q^*); \ \phi = [g(Q^*) - \xi \cdot Q^*]/g(Q^*)^2;$$

$$\begin{aligned}
\eta &= h(Q^*) + h'(Q^*) \cdot Q^*; \\
a &= \lambda\beta_t\xi N^{-1};\ b = \lambda\beta_c\eta N^{-1}; \\
B &= \phi^2 + \kappa^2 - a^2 - 2b;\ D = (\phi\kappa + b)^2 - a^2\kappa^2; \\
\Omega &= (-B + \sqrt{B^2 - 4D})/2.
\end{aligned}$$

If $\Omega \notin \mathbb{R}_{++}$, then the equilibrium solution 2.8 of system 2.7 is stable for all $\Delta > 0$; Otherwise, it is stable if $\Delta < \Delta_{cr}$ and unstable if $\Delta \geq \Delta_{cr}$, where

$$\Delta_{cr} = \frac{1}{\sqrt{\Omega}} \arccos\left(-\frac{b\kappa}{a(\kappa^2 + \Omega)} - \frac{\phi}{a}\right). \tag{2.9}$$

If there is no air pollution information, we can assume that motorists have no perceived cost for exposure to air pollution, thus $\beta_c = 0$. Therefore, the stability condition in this case reduces to the following: If $\beta_t \leq \lambda^{-1}\xi^{-1}\phi N$, the equilibrium solution 2.8 is always stable; otherwise, we obtain the critical delay by substituting $b = \lambda\beta_c\eta_N^{-1} = 0$ in equation (2.9). This enables us to rigorously compare the stability implications for the cases with and without posting of air pollution information.

Figure 2.4 shows the behavior of the system under the same input and initial conditions as the example in Figure 2.3 with delay $\Delta = 4$ (min), except that we set $\beta_c = 0$ to represent the case without posting of air pollution information. We can see that the traffic queues oscillate with a similar magnitude but at a greater frequency than in the case with air pollution information posted ($\beta_c = 0.2$ (\$/(μg/m^3))) and the longer delay, $\Delta = 10$ (min) (see the lower plot in Figure 2.3). The critical delay for the case without air pollution information is $\Delta_{cr} = 3.533$ (min), which is 35% shorter than for the case with air pollution information. This implies that disclosure of air pollution information can significantly help improve traffic stability: without posting of air pollution information, the timeliness requirement of travel time posting is much more stringent.

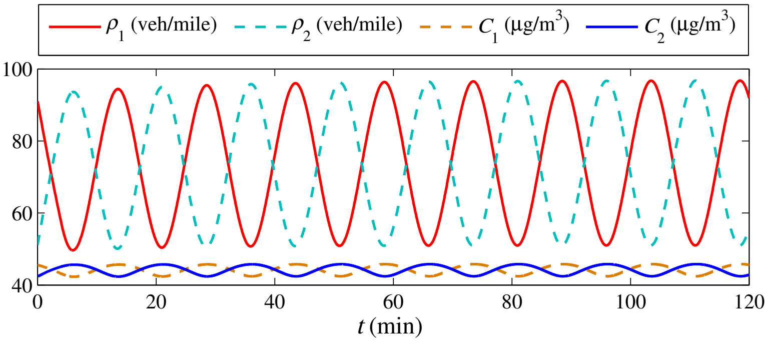


Figure 2.4: Example: $\lambda = 30,\ \beta_c = 0,\ \Delta = 4$.

Actually such a stability-enhancing effect is guaranteed under a mild assumption on the total emission source function, as stated in our second main theorem. For convenience in stating the results, we denote Δ_{cr}^{yes} and Δ_{cr}^{no} as the critical delays (defined in equation (2.9)) for the cases with ($\beta_c > 0$) and without ($\beta_c = 0$) posting air pollution information, respectively. The proof of this result is also provided in the Appendix A.

Theorem 2.2 *If the total emission source strength (as a function of queue length x defined in equation (2.3)) $S(x) = h(x) \cdot x$ is strictly increasing at the equilibrium solution Q^* defined in (2.8), i.e.*

$$\frac{dS(x)}{dx} = \frac{d(h(x)\cdot x)}{dx}\Big|_{x=Q^*} > 0, \tag{2.10}$$

then we have: 1) if $\Delta_{cr} < \infty$, then Δ_{cr} is strictly increasing in $\beta_c \geq 0$; 2) if $\Delta_{cr}^{\text{no}} = \infty$, then $\Delta_{cr}^{\text{yes}} = \infty$.

Theorem 2.2 has the following important implications: 1) When condition (2.10) holds, the critical delay threshold (if it is finite) under the provision of air pollution information is always longer (i.e., less stringent for the timeliness of travel time estimation/posting) than otherwise; 2) If the users attach more value to the air pollution (i.e., health implications) in their routing decision, the critical delay threshold becomes longer, i.e., the system can tolerate even extended time lag in posting travel time information to still achieve stable traffic; Moreover, 3) For situations where the system is always stable (regardless of how long the information delay Δ is) without provision of air pollution information, it remains stable when air pollution information is added.

Notice that the total emission source strength $S(Q) = h(Q) \cdot Q$ is generally higher when there are more vehicles on the link (i.e., Q is bigger) (see for example, Figure 2.13 in Appendix B). Hence the function $S(Q)$ is in general strictly increasing in Q at the equilibrium solution Q^*. Therefore, the sufficient condition (2.10) holds in general and is a natural result of the typical relationship between total emission source strength and the number of vehicles on the link. Theorem 2.2 is therefore quite general, indicating that informing drivers of air pollution information can improve traffic stability and mitigate congestion effectively in general cases, especially when the provision of travel time information is hindered by delays.

2.3.2 Simulation results

To verify the phenomenon predicted by our analytical model, we conducted a stochastic simulation of the east-bound morning traffic (7:00–8:00 AM) on the George Washington Bridge, which is modeled by two links: one for the upper level (with 4 lanes) and the other for the lower level (with 3 lanes). Traffic flow was simulated by the cellular automaton model [92], which is a commonly-used microscopic traffic simulation model that can reproduce key features of traffic flow such as stop-and-go conditions [92]. In each simulation interval, information on the travel time estimate and $PM_{2.5}$ concentration was fed back to the arriving driver, who then chose either the upper level or the lower level with probabilities computed by equation (2.5). Travel time information was delayed by 5 minutes. We use $PM_{2.5}$ for our experiment since its hourly average background concentration data for the GWB area is available [96]. If we could get data for other pollutants such as NO_x, we can easily do the same experiment by using new C_b, κ and function h, but the qualitative results will stay consistent, as we discussed in Remark 2.1. In order to simulate pollutant level in a discrete time setting, $PM_{2.5}$ concentration was modeled by discretizing the second equation in (2.7), and an empirical function $h(Q)$ was used. Stochasticity in our simulation included: 1) vehicle arrivals following a Poisson process, 2) randomness in vehicle deceleration [92], and 3) white noise in the air pollution concentration to represent modeling errors.

Figure 2.5 shows summary statistics on the total numbers of vehicles and $PM_{2.5}$ concentrations on link 1 and link 2, respectively. In the case without the disclosure of air pollution information, numbers of vehicles and $PM_{2.5}$ concentrations on the two links show persistent oscillation. This unstable pattern is statistically significant and very similar to the results derived from the analytical model. In contrast, both numbers of vehicles and $PM_{2.5}$ concentrations become stable after a short period of variation when $PM_{2.5}$ concentration information is provided. This result verifies the traffic-stabilizing effect of disclosure of air pollution information. Note that in this example the variation of $PM_{2.5}$ concentration is quite small (within $2\mu g/m^3$), given the data we used for the simulation. However, the concentration level can vary more significantly if a different pollutant or emission rate function is used such that emission rate changes more dramatically as vehicle speed changes. We mainly focus on the qualitative observations through our example (see Remark 2.1).

We also observe that without air pollution information, there is an ascending trend in the numbers of vehicles on account of accumulation of vehicles waiting to enter the links, but this does not occur when air pollution information is posted. The reason is that with improved traffic stability across links, higher demand can be accommodated due to the decline in the frequency of service rate reductions caused by traffic oscillation. This is an interesting observation, as each link only had a finite number of servers.

In addition, Figure 2.6 shows the aggregate results of average travel time and choice probabilities of the two levels on the GWB over 10^3 simulation runs. We observe oscillations of travel time and link choice probabilities, which again reflect unstable traffic pattern under no provision of air pollution information. In contrast, the travel time and choice probabilities stay quite stable under provision of $PM_{2.5}$ concentration information, with only notable variations in the beginning. Note that because the lower lever of GWG has one less lane (smaller supply) compared to the upper level, choice probability is generally higher for the upper level. In addition, from Figure 2.5 we made an observation that when no air pollution information is posted the number of cars on two links have a gradual increasing trend due to accumulation of waiting vehicles. Accordingly, here in 2.6 we see that when there is no air pollution information, the travel time also rises as time goes on, which again reflects the formation of upstream congestion that validates the simulation model. Figure 2.7 further plots the total number of vehicles together with the number of vehicles traveling on the lanes, we can see a clear increasing gap between the two curves, which indicates an accumulation of vehicles waiting to enter the lanes. And there are more vehicles waiting to enter the lanes at the lower level since it has fewer lanes (thus fewer number of servers). Posting real-time air pollution information, however, prevents this phenomenon by maintaining more balanced service rates across links that serve the demand better.

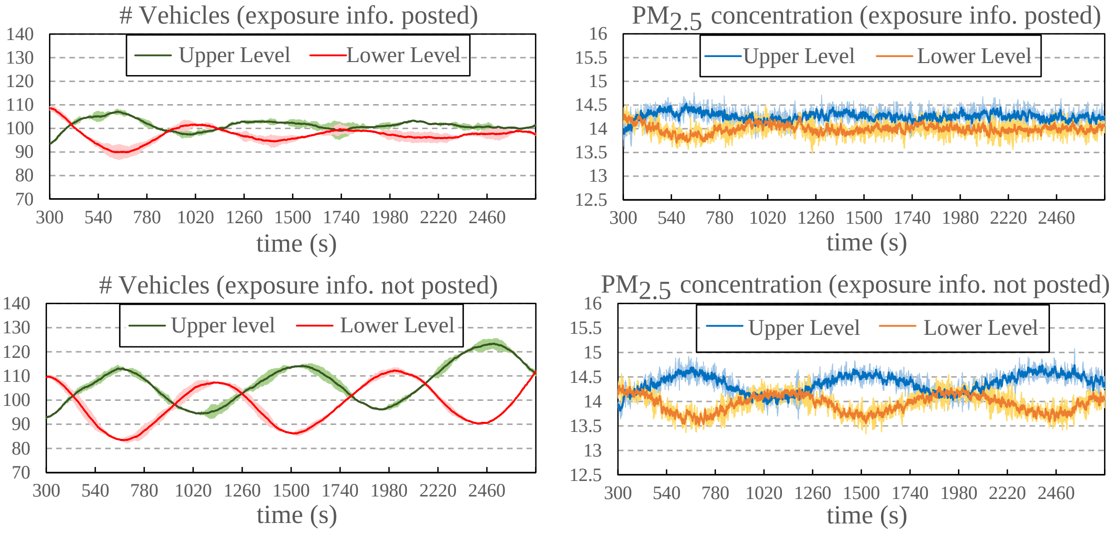


Figure 2.5: Numbers of vehicles and $PM_{2.5}$ concentrations ($\mu g/m^3$) on the two links: air pollution information posted (upper two plots) versus not posted (lower two plots). The dark center line on each curve represents the average of 10^3 independent simulation runs; the lighter-colored area surrounding the dark line on each curve represents the 95% confidence interval for the mean estimate.

Moreover, in Table 2.2 we report the average numbers of waiting vehicles outside the links (AWV) and the total numbers of vehicle-stops (TVS) during the simulation period of 300~2700 s. These measures again indicate that disclosure of air pollution information helps mitigate traffic congestion. Additional details on the simulation data and model as well as results under different delay and

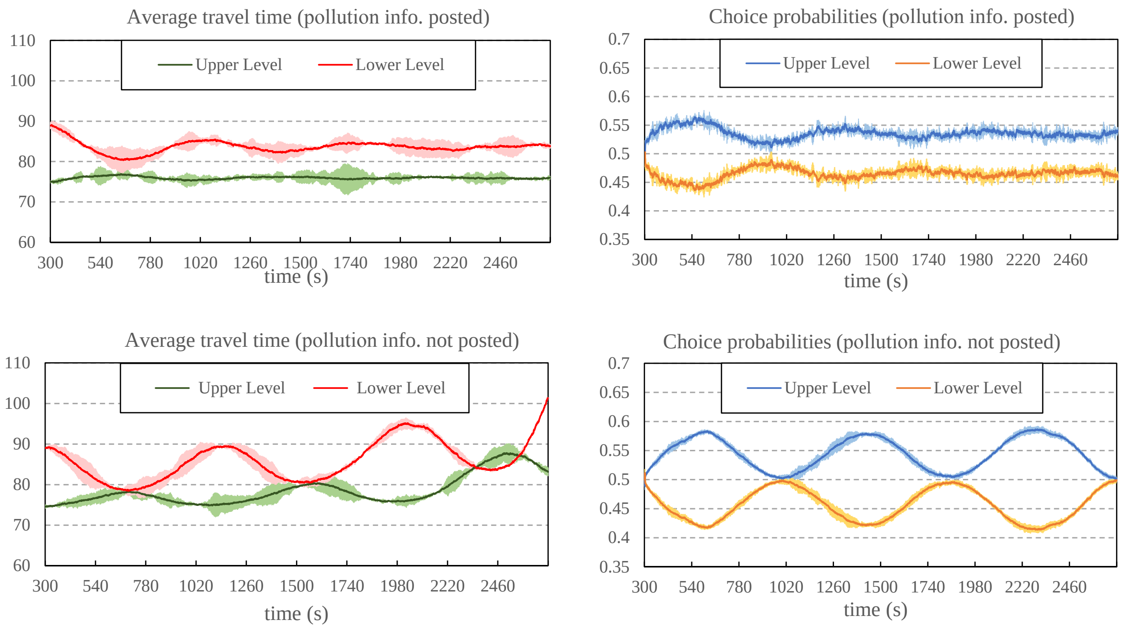


Figure 2.6: Average travel time (s) and choice probabilities of two links: air pollution information posted (upper two plots) versus not posted (lower two plots). The dark center line on each curve represents the average of 10^3 independent simulation runs; the lighter-colored area surrounding the dark line on each curve represents the 95% confidence interval for the mean estimate.

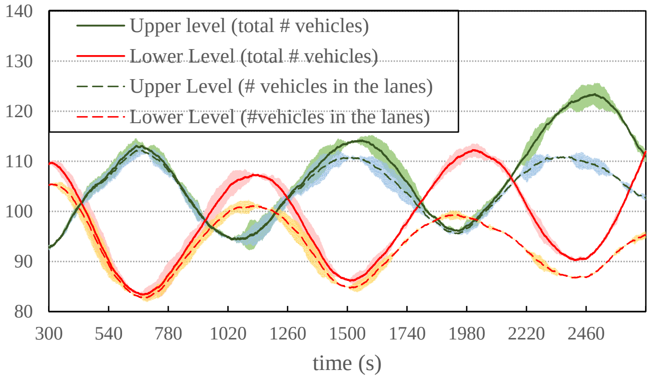


Figure 2.7: Total number of vehicles (including those waiting) and number of vehicles within the lanes when air pollution information is not posted. The dark center line on each curve represents the average of 10^3 independent simulation runs; the lighter-colored area surrounding the dark line on each curve represents the 95% confidence interval for the mean estimate.

WTPs are provided in the Appendix B.

Table 2.2: Other outputs (mean [standard deviation] of 10^3 samples)

Scenario	AWV	TVS
$\text{PM}_{2.5}$ info. posted	2.6 [0.590]	13730.4 [259.5]
$\text{PM}_{2.5}$ info. not posted	8.6 [0.493]	16562.9 [173.2]

2.4 Discussion

By Theorem 2.2 as well as our numerical and stochastic simulation results, we have shown that disclosure of air pollution information to travelers enhances traffic stability and mitigates congestion in simple transportation networks. Now we will examine such benefits systematically under variations of the system parameters. By Theorem 2.1, traffic stability is affected by a number of factors: 1) the total arrival rate λ and the number of links N; 2) motorists' WTPs β_t and β_c; and 3) the emission source strength and the dispersion parameters. Via numerical analysis we will discuss the quantitative effects of these three groups of parameters on system performance. For analysis of the sensitivity to the specific model parameters of concern, we keep the other parameters fixed at the values used in the example in Figure 2.3. Empirical insights, policy implications, and practical recommendations are also discussed.

2.4.1 Total arrival rate λ and number of links N

Figure 2.8 shows a plot of the critical delay Δ_{cr} for various arrival rate λ, and for both $N = 2$ (two links) and $N = 3$ (three links). We can see that when λ is low (i.e., during the off-peak period), traffic across links always converges to the equilibrium. However, as λ increases (to 26 or higher when $N = 2$, and to 39 or higher when $N = 3$), the queues remain asynchronous when information delay Δ reaches or exceeds the critical level, Δ_{cr}. The critical delay Δ_{cr} decreases with λ, but at a decreasing rate. This implies that when the arrival rate is relatively high (i.e., during peak hours), the improvement in system stability which stems from the disclosure of air pollution information can be substantial and consistent, even over a wide range of arrival rates (e.g., compare Figure 2.4 to the upper plot in Figure 2.3). If we increase the number of links from 2 to 3, the system is always stable under the same range of λ as in Figure 2.8. This implies that under fixed total demand and information delay, the larger the number of links, the more stable the traffic network. In addition, the critical delay Δ_{cr} stays constant if λ/N is held fixed. This invariance property can be deduced from Theorem 2.1 (see Appendix A). Therefore, making air pollution information available on a larger number of links can help achieve greater stability in terms of the traffic distribution.

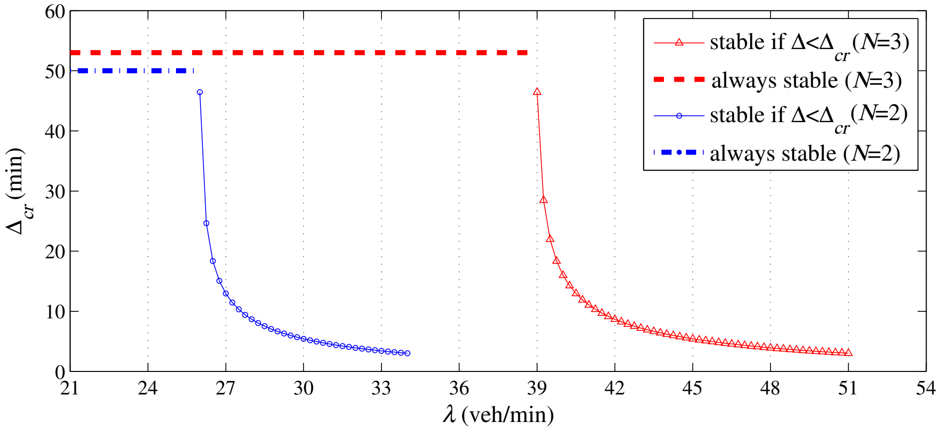


Figure 2.8: Stability region and critical delay for $N = 2$ and $N = 3$, under various total arrival rate λ.

2.4.2 Behavior parameters β_t and β_c

Figure 2.9 shows a plot of the critical delay as a function of the WTP for travel time, β_t, and the WTP for air pollution, β_c. Each blue curve displays the value of Δ_{cr} (if $\Omega > 0$) as a function of β_t (left plot) or β_c (right plot), with the other model parameters fixed. Each dashed red line corresponds to the case $\Omega \leq 0$ when the system is stable for all $\Delta > 0$. The blue curves show that Δ_{cr} is monotonically decreasing in β_t (which is proved rigorously in the Appendix A for the case where $g'(Q^*) > 0$) and at a decreasing rate, while Δ_{cr} is monotonically increasing in β_c (consistent with Theorem 2.2) and at an increasing rate. The lengths of the dashed red lines illustrate that the width of the "always stable" region is increasing in β_c and decreasing in β_t. These observations imply that traffic stability improves as β_t decreases or as β_c increases. Therefore, a key conclusion from these comparisons is that the traffic distribution in the network can achieve a stable equilibrium without oscillation provided that the marginal rate of substitution β_c/β_t ((μg/m^3)/min) is sufficiently large, that is, that motorists attach a certain value to the perceived impact of air pollution compared to that of travel time. For example, when β_t increases to 0.6 (\$/min), the traffic system can attain stability if the WTP for air pollution, β_c, reaches 0.32 (μg/m^3) (about half of β_t), even when the travel time information delay is as long as 10 minutes. Note also that if there is no air pollution information posted, motorists will make their routing decisions unaware of the level of air pollution ($\beta_c = 0$). In this case, a delay in the travel time information as short as $\Delta = 6$ (min) will cause unstable queues for any $\beta_t \geq 0.36$ (See the lowest blue curve in the left plot in Figure 2.9). Again, these observations show the substantial value of disclosure of air pollution information in the stabilization of traffic.

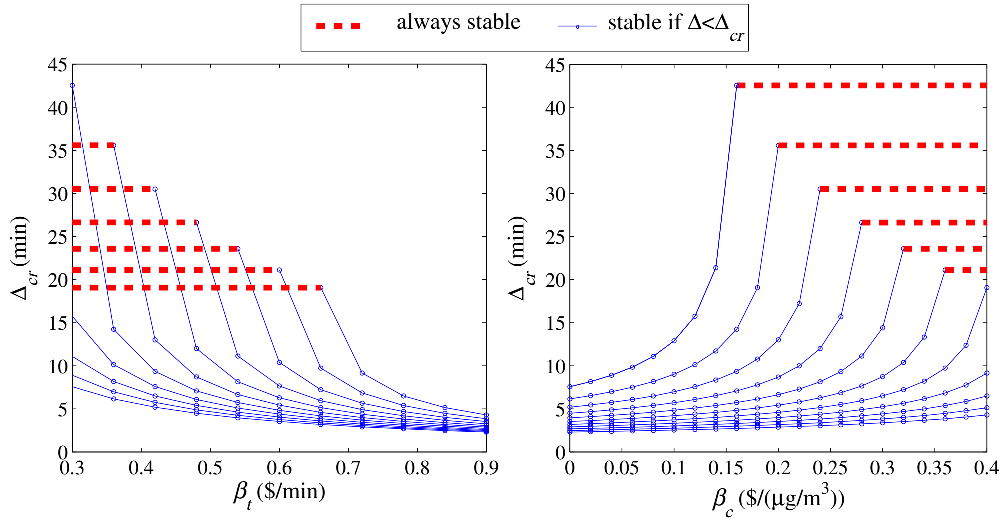


Figure 2.9: Stability region and critical delay under various β_t and β_c. Each curve in the left plot shows Δ_{cr} as a function of β_t at a fixed β_c (from $\beta_c = 0$ to 0.4 from bottom to top, in increments of 0.04); each curve on the right plot shows Δ_{cr} as a function of β_c at a fixed β_t (from $\beta_t = 0.3$ to 0.9 from top to bottom, in increments of 0.06).

2.4.3 Emission strength parameter c and dispersion factor κ

In the left 3D plot in Figure 2.10, we show the critical delay Δ_{cr} as a function of the dilution rate constant κ and the coefficient c in the average emission source strength function: $h(Q) = c \cdot \exp(-0.0058Q)$. Note that given the number of vehicles on a link, c is proportional to the average emission strength per vehicle. We can see that with the other parameters held fixed, the dependence of the stability of the traffic density on these emission and dispersion parameters is not monotone. When the dilution rate is very low ($\kappa \leq 0.5$), the traffic can attain a stable condition and endure a larger Δ_{cr} under combinations of larger κ and smaller c. When the dilution rate increases to $\kappa > 0.5$, Δ_{cr} surges as soon as the average emission strength per vehicle exceeds 0.4. In particular, traffic can always attain stability when $0.5 < \kappa < 3c$ (i.e., the triangular ceiling in the 3D mesh). Δ_{cr} drops dramatically when κ exceeds $3c$, but the rate of this reduction decreases thereafter. The rate of the decrease is higher for large values of c. In this region, the system also becomes slightly more stable as the emission strength parameter c increases. These observations imply that under relatively low dispersion intensity (e.g., low wind speed, stable atmospheric conditions), which is conducive to accumulation of air pollution, a larger emission source factor c can lead to more effective stability improving effect of disclosure of air pollution information. When the emission strength parameter c is not very high and the dispersion intensity is relatively high, posting of air pollution information has an almost constant positive effect in terms of enhancing traffic stability. The "always stable" situation is achieved at moderate to high dilution rate and large emission strength factor, as reflected in the triangular region (roughly $0.5 < \kappa < 3c$). It is in this region of the emission–dispersion space that posting of air pollution information yields the greatest gain in traffic stability. The explanations are intuitive: 1) Under such emission and dispersion conditions, the on-road pollutant concentration can notably reflect different queue lengths for different links, thereby affecting travelers' route choice, which in turn rebalances traffic (this can be seen from the equilibrium pollution concentration in solution 2.8, as κ appears in the denominator and h is proportional to c); 2) with relatively larger κ, vehicle emissions are dispersed and diluted fast enough, thus near-source measurement of on-road air pollution gives a more accurate real-time indicator of traffic densities than does the posted estimate of the travel time, which usually suffers from a time lag.

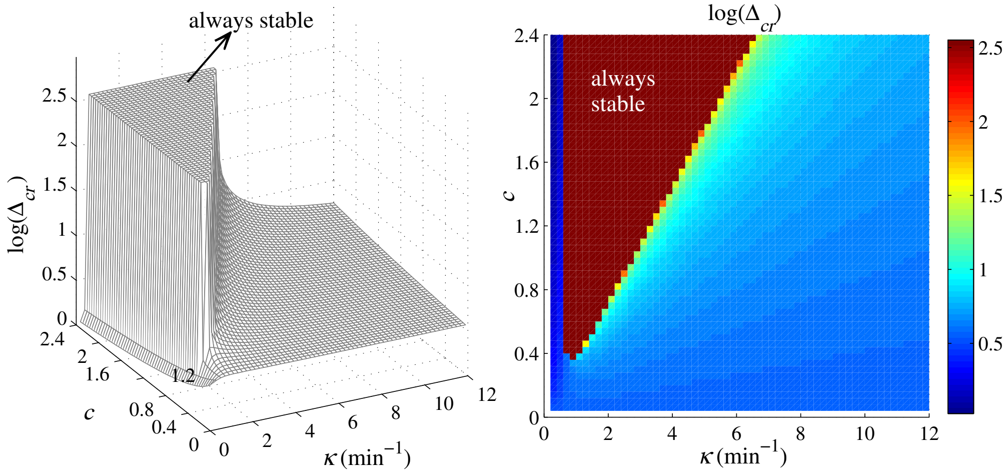


Figure 2.10: Stability region and critical delay under various values of the emission strength coefficient c and various dilution rates κ. The 3D mesh (left) shows the log-scaled Δ_{cr} as a function of both c and κ, its flat roof corresponds to the “always stable” region. The 2D plot (right) is a top view of the 3D mesh; the color indicates the value of the critical delay Δ_{cr} in log scale.

2.5 Conclusions

This study analyses how disclosing on-road air pollution information to drivers may help improve traffic stability for proactive congestion management. Our theoretical model shows that there can be notable benefits gained by this approach to traffic control, especially when the timeliness of travel time reporting is limited. Sensitivity analysis of demand, behavior, and environmental parameters indicates that disclosure of air pollution information has a robust stability-enhancing effect on traffic. In particular, such effect is observed even if the travelers have relatively low valuation on the air pollution information, and the effect is most evident when the pollution dispersion intensity is moderate to high and vehicle emission strength factor is large. Simulation of morning peak-period traffic on the George Washington Bridge shows that posting real-time air pollution concentration on each level of the bridge results in smoother and more stable traffic. Our findings indicate that disclosure of air pollution information, which affects travelers' routing behavior, can be an effective tool for alleviating congestion.

For future research, we need to investigate the effect of the proposed strategy more comprehensively using driving simulator or a real-world experiment. Study on the related sensing method and risk communication approaches are also necessary for improving the efficacy of the proposed strategy and bring it closer to real-world application.

2.6 Appendix A: Proofs of the Main Theorems, Related Results

In this last section we provide the proofs of the key analytical results, including the stability condition (Theorem 2.1) and some related results, as well as a sufficient condition of the stability enhancing effect of posting air pollution information (Theorem 2.2).

2.6.1 Proof of Theorem 2.1

Proof: To understand the stability of (2.7) near the equilibrium solution (2.8), we first add perturbations u_i and w_i ($i = 1, ..., N$) to the equilibrium solution (2.8)

$$\begin{aligned} Q_i(t) &= Q^* + u_i(t),\ i = 1, ..., N; \\ C_i(t) &= \frac{Q^* h(Q^*)}{\kappa} + C_b + w_i(t),\ i = 1, ..., N. \end{aligned} \tag{2.11}$$

Substituting (2.11) into system (2.7), we get

$$\begin{aligned}
\dot{u}_i(t) &= \lambda \frac{\exp[-\beta_t g(Q^* + u_i(t-\Delta)) - \beta_c C_i(t)]}{\sum_{j=1}^{N} \exp[-\beta_t g(Q^* + u_j(t-\Delta)) - \beta_c C_j(t)]} - \frac{Q_i(t)}{g(Q_i(t))},\ i = 1, ..., N; \\
\dot{w}_i(t) &= h(Q^* + u_i(t)) \cdot (Q^* + u_i(t)) - h(Q^*) \cdot Q^* - \kappa w_i(t),\ i = 1, ..., N.
\end{aligned} \tag{2.12}$$

Performing a Taylor expansion on the RHS of (2.12) around point $u_i(t) = w_i(t) = 0$ $(i = 1, ..., N)$, we obtain the linearized version of (2.12) as

$$\begin{aligned}
\dot{u}_i(t) &= -\frac{\lambda}{N^2}[\beta_t g'(Q^*)((N-1)u_i(t-\Delta) - \sum_{j\neq i} u_j(t-\Delta)) \\
&+ \beta_c((N-1)w_i(t-\Delta) - \sum_{j\neq i} w_j(t))] - \frac{g(Q^*) - g'(Q^*)Q^*}{g(Q^*)^2} u_i(t),\ i = 1, ..., N; \\
\dot{w}_i(t) &= [h'(Q^*) \cdot Q^* + h(Q^*)]\, u_i(t) - \kappa w_i(t), \qquad i = 1, ..., N.
\end{aligned} \tag{2.13}$$

Note that analyzing the stability of system (2.7) is equivalent to analyzing the stability of the linearized system (2.13), this is based on non-trivial results in analysis of nonlinear dynamical systems [59, 116]. Although system (2.13) is linear, it includes many equations that need to be analyzed. However, we will show that through a series of transformations we can simplify the analysis of these $2N$ equations to 2 equations. To this end, we apply two groups of transformations $\{v_i(t)\}$ and $\{x_i(t)\}$ defined as

$$\begin{aligned}
v_1(t) &= \sum_{i=1}^{N} u_i(t); \\
v_2(t) &= u_1(t) - u_2(t); \\
&... \\
v_N(t) &= u_{N-1}(t) - u_N(t); \\
x_1(t) &= \sum_{i=1}^{N} w_i(t); \\
x_2(t) &= w_1(t) - w_2(t); \\
&... \\
x_N(t) &= w_{N-1}(t) - w_N(t),
\end{aligned}$$

and the change of variables

$$\begin{aligned}
\xi &= g'(Q^*);\\
\phi &= \frac{1}{g(Q^*)} - \frac{\xi Q^*}{g(Q^*)^2};\\
\eta &= h(Q^*) + h'(Q^*) \cdot Q^*. \qquad (2.14)
\end{aligned}$$

Then system (2.13) is equivalent to

$$\begin{aligned}
\dot{v}_1(t) &= -\phi v_1(t);\\
\dot{v}_2(t) &= -\frac{\lambda}{N}(\xi\beta_t v_2(t-\Delta) + \beta_c x_2(t)) - \phi v_2(t);\\
&\ldots\\
\dot{v}_N(t) &= -\frac{\lambda}{N}(\xi\beta_t v_N(t-\Delta) + \beta_c x_N(t)) - \phi v_N(t);\\
\dot{x}_1(t) &= \eta v_1(t) - \kappa x_1(t);\\
\dot{x}_2(t) &= \eta v_2(t) - \kappa x_2(t);\\
&\ldots\\
\dot{x}_N(t) &= \eta v_N(t) - \kappa x_N(t). \qquad (2.15)
\end{aligned}$$

If we pair up the i^{th} equation and the $(N+i)^{\text{th}}$ equation ($i = 1, ..., N$) in (2.15), we notice that only two variables v_i and x_i are involved in these two equations. Therefore, this transformation decouples the variables originally involved in system (2.13). This simplifies the analysis significantly.

The solution to the first equation in (2.15) is $v_1 = c_1 \exp(-\phi t)$ with constant c_1, so it is stable since $\phi = q'(Q^*) > 0$. Then we know that the $(N+1)^{\text{th}}$ equation in (2.15) is also stable since $\kappa > 0$ and η is finite. To analyze the rest $2(N-1)$ equations in (2.15), we notice that the paired up i^{th} and $(N+i)^{\text{th}}$ equations in (2.15) have the same form in terms of variables x_i and v_i for all $i = 2, ..., N$, hence we only need to analyze one such pair. Now to explore the stability of equation

i and equation $(N+i)$, $i \in \{2, ..., N\}$, we can substitute the following exponential expressions

$$\begin{aligned} v_i(t) &= a_i \exp(rt),\ i = 2, ..., N; \\ x_i(t) &= b_i \exp(rt),\ i = 2, ..., N. \end{aligned} \tag{2.16}$$

Thus, for each $i = 2, ..., N$, substituting (2.16) into the i^{th} and the $(N+i)^{\text{th}}$ equations in (2.15), we get

$$\begin{aligned} a_i r &= -\frac{\lambda}{N}[\xi\beta_t a_i \exp(-r\Delta) + \beta_c b_i] - \phi a_i,\ i = 2, ..., N; \\ b_i r &= \eta a_i - \kappa b_i, \quad i = 2, ..., N. \end{aligned} \tag{2.17}$$

Using the second equation in (2.17), we solve for a_i in terms of b_i and obtain for $i = 2, ..., N$

$$a_i = \frac{\kappa + r}{\eta} b_i,$$

then substitute this back to (2.17), we get

$$\frac{(\kappa + r)(\phi + r)}{\eta} = -\frac{\lambda}{N}\left[\frac{\xi\beta_t(\kappa + r)}{\eta}\exp(-r\Delta) + \beta_c\right]. \tag{2.18}$$

With the transcendental equations for parameter r, it only remains for us to find the transition between stable and unstable solutions. Characteristic equations of the form (2.17) are often studied in order to understand changes in the local stability of equilibria of delay differential equations. Thus, it is important to determine the values of the delay at which there are roots with zero real part. When parameter r crosses the imaginary axis, the stability of the equilibrium changes. In the fully non-linear system this transition generally occurs in a Hopf bifurcation, in which a pair of roots crosses the imaginary axis and a limit cycle occurs. Thus, to find the critical delay for the change of stability, we set

$r = i\omega,\ \omega \in \mathbb{R}$ and substitute this in (2.18). This gives

$$\exp(-i\omega) = [\cos(\omega\Delta) - i\sin(\omega\Delta)] = -\frac{(\phi + i\omega)N}{\lambda\xi\beta_t} - \frac{\beta_c\eta}{\xi\beta_c(\kappa + i\omega)},$$

which is equivalent to

$$-\lambda\xi\beta_t(\kappa + i\omega)[\cos(\omega\Delta) - i\sin(\omega\Delta)] = (\phi + i\omega)(\kappa + i\omega)N + \lambda\beta_c\eta. \tag{2.19}$$

Setting the real and imaginary parts for Equation (2.19) to zero, we obtain the following system of equations

$$\begin{aligned} -\lambda\beta_c\eta + (\omega^2 - \phi\kappa)N &= \lambda\xi\beta_t\kappa\cos(\omega\Delta) + \lambda\xi\beta_t\omega\sin(\omega\Delta); \\ N(\phi + \kappa)\omega &= -\lambda\xi\beta_t\omega\cos(\omega\Delta) + \lambda\xi\beta_t\kappa\sin(\omega\Delta). \end{aligned} \tag{2.20}$$

Solving (2.20) for $\sin(\omega\Delta)$ and $\cos(\omega\Delta)$, we obtain

$$\begin{aligned} \sin(\omega\Delta) &= \frac{-\lambda\beta_c\eta\omega + N\omega(\kappa^2 + \omega^2)}{\lambda\xi\beta_t(\kappa^2 + \omega^2)}; \\ \cos(\omega\Delta) &= \frac{-\lambda\beta_c\eta\kappa - N\phi(\kappa^2 + \omega^2)}{\lambda\xi\beta_t(\kappa^2 + \omega^2)}. \end{aligned} \tag{2.21}$$

So based on (2.21), using the equation $\sin(\omega\Delta)^2 + \cos(\omega\Delta)^2 = 1$, we have

$$\lambda^2\xi^2\beta_t^2(\kappa^2 + \omega^2)^2 = [-\lambda\beta_c\eta\omega + N\omega(\kappa^2 + \omega^2)]^2 + [\lambda\beta_c\eta\kappa + N\phi(\kappa^2 + \omega^2)]^2,$$

we can factor out a term $(\kappa^2 + \omega^2)$ on the RHS and cancel that of the LHS and through rearrangement. This yields an expression that sets a 4^{th} order polynomial of ω to 0

$$\omega^4 + B\omega^2 + D = 0. \tag{2.22}$$

The expressions for coefficients B and D in (2.22) are

$$\begin{aligned} B &= \kappa^2 + \phi^2 - a^2 - 2b; \\ D &= (\phi\kappa + b)^2 - a^2\kappa^2, \end{aligned}$$

where we define

$$\begin{aligned} a &= \lambda\beta_t\xi N^{-1}; \\ b &= \lambda\beta_c\eta N^{-1}. \end{aligned} \tag{2.23}$$

Since we have $\omega^2 \geq 0$, Equation (2.22) has real solutions if only if $\mathcal{W}_0 \neq \emptyset$, where the set

$$\mathcal{W}_0 := \mathbb{R}_+ \cap \left\{ \frac{-B \pm \sqrt{B^2 - 4D}}{2} \right\}.$$

Now we also claim that $\omega = 0$ cannot be a solution to (2.21). This is because when $\omega = 0$, we have a contradiction for the second equation in (2.21), since Δ is finite and $\eta, \phi > 0$,

$$\cos(0) = 1 = \frac{-\lambda\beta_c\eta\kappa - N\phi\kappa^2}{\lambda\xi\beta_t\kappa^2} < 0.$$

It hence follows that (2.21) has real solution of ω if only if the set

$$\mathcal{W} := \mathbb{R}_{++} \cap \left\{ \frac{-B \pm \sqrt{B^2 - 4D}}{2} \right\} \neq \emptyset.$$

Thus if we have $\mathcal{W} \neq \emptyset$, let $\Omega \in \mathcal{W}$, by substituting it into the second expression of (2.21) we have the following due to symmetry of the cosine function,

$$\begin{aligned} \cos(\sqrt{\Omega}\Delta) &= \frac{-\lambda\beta_c\eta\kappa - N\phi(\kappa^2 + \Omega)}{\lambda\xi\beta_t(\kappa^2 + \Omega)} \\ &= -\frac{b\kappa}{a(\kappa^2 + \Omega)} - \frac{\phi}{a}. \end{aligned} \tag{2.24}$$

Note that there are infinite number of solutions of Δ to the transcendental equation (2.24) due to periodicity of cosine function. However, we are interested in the smallest solution to (2.24) which we call it the critical delay: Δ_{cr}. Δ_{cr} determines the stability region of the equilibrium (2.8) due to Lemmas 1 & 2 that we will prove later.

Note that for any $\Omega \in \mathcal{W}$, the smallest solution of (2.24) is

$$\Delta(\Omega) = \frac{1}{\sqrt{\Omega}} \arccos\left(-\frac{b\kappa}{a(\kappa^2 + \Omega)} - \frac{\phi}{a}\right). \tag{2.25}$$

Now a key observation is that $\Delta(\Omega)$ defined in (2.25) is decreasing in Ω, this is because both

$$\frac{1}{\sqrt{\Omega}} \quad \text{and} \quad \arccos\left(-\frac{b\kappa}{a(\kappa^2 + \Omega)} - \frac{\phi}{a}\right)$$

are non-negative and decreasing in Ω.

Hence when $\mathcal{W} \neq \emptyset$, we want to find $\Omega^* = \max \mathcal{W}$. Note that if $\mathcal{W} \neq \emptyset$, then it must be true that $\sqrt{B^2 - 4D} \geq 0$, it follows then that $-B + \sqrt{B^2 - 4D} \geq -B - \sqrt{B^2 - 4D}$, so we can deduce

$$\Omega^* = \max \mathcal{W} = \frac{-B + \sqrt{B^2 - 4D}}{2}. \tag{2.26}$$

And clearly, we have $\mathcal{W} \neq \emptyset \Leftrightarrow \Omega^* \in \mathbb{R}_{++}$.

Then the smallest solution $\Delta = \Delta_{cr}$ to Equation (2.24) is

$$\Delta_{cr} = \frac{1}{\sqrt{\Omega^*}} \arccos\left(-\frac{b\kappa}{a(\kappa^2 + \Omega^*)} - \frac{\phi}{a}\right). \tag{2.27}$$

Therefore, by Lemmas 1 & 2 (see Section 2.1), we conclude that if $\Omega^* \in \mathbb{R}_{++}$, then the equilibrium (2.8) of system (2.7) is stable if only if $\Delta < \Delta_{cr}$; Otherwise ($\Omega^* \notin \mathbb{R}_{++}$), the equilibrium (2.8) is always stable (Ω^* and Δ_{cr} are defined in (2.26) and (2.27), respectively). ■

2.6.2 Related results of Theorem 2.1

Two lemmas needed for Theorem 2.1

One important part for the proof of Theorem 2.1 is to show that whenever the time delay Δ in travel time provision exceeds the smallest solution to the transcendental equation (2.24) (which we defined as Δ_{cr} in (2.27)), the system will stay unstable. In other words, as Δ increases from 0 to ∞, the real part of the eigenvalues of the linearized system switches sign only once (from negative to positive) when Δ crosses Δ_{cr}.

For example, Figure 2.11 shows different solutions Δ to the transcendental equation (2.24) under various total arrival rate λ (veh/min) given other parameters fixed. We call these curves "Hopf curves" [113, 116], each of which depicts one system parameter (Δ here) as a function of another parameter (λ here) when there exists a pair of pure imaginary eigenvalues of a dynamical system ((2.7) here). We need to show that under a given λ, if Δ increases to Δ_{cr} (the critical dalry calculated by (2.27), which is on the lower left curve in Figure 2.11), the system changes from stable to unstable and remains unstable as Δ increases.

To show this, we use perturbation techniques. Let $\Delta_1, \Delta_2, ..., \Delta_\infty$ be the solutions of the transcendental equation (2.24) (in particular $\Delta_1 = \Delta_{cr}$). Consider the i^{th} solution Δ_i for any $i = 1, 2, ...$, suppose that we make a small perturbation on the order of $\epsilon\Delta_p$ ($\epsilon \ll 1$), i.e.,

$$\Delta = \Delta_i + \epsilon\Delta_p. \tag{2.28}$$

Then the root r of Equation (2.17) corresponding to Δ_i will be also slightly perturbed from the pure imaginary value it would take at $\Delta = \Delta_i$, we can write

$$r = i\omega + \epsilon(ix + y), \tag{2.29}$$

where $x \neq 0$ and $y \neq 0$ denote the imaginary and real parts of the perturbation, which can be determined in terms of Δ_i and Δ_p.

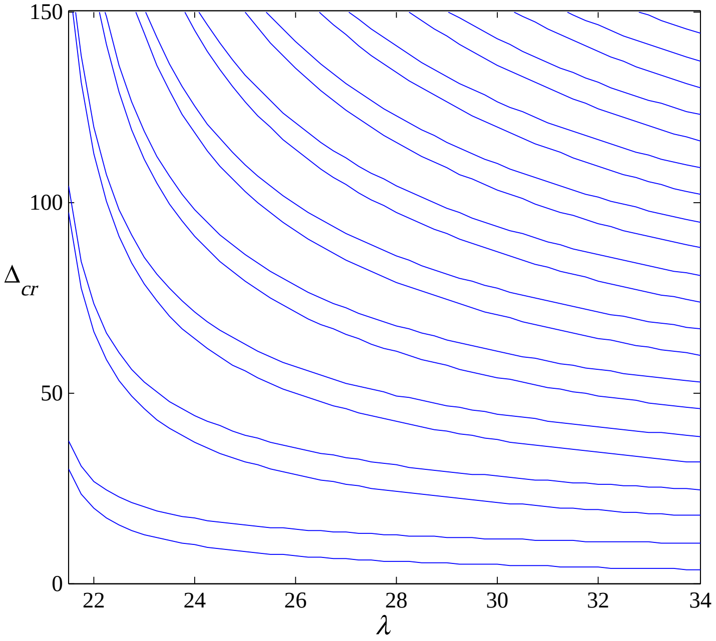


Figure 2.11: Hopf curves of delay Δ and total arrival rate λ (given other parameters: $N = 2$, $\beta_t = 0.5$, $\beta_c = 0.1$, $\kappa = 12$, $g(x) = 0.0526x + 1.0$, $h(x) = 1.176e^{-0.0058x}$).

We are mainly interested in the real part y, whose sign determines the stability of the system. The following two key Lemmas are related to this analysis.

Lemma 2.1 *Under the perturbation formula (2.28) for Δ_i and (2.29) for the corresponding root r, the real part of the perturbed r is*

$$y = \frac{\Delta_p \omega^2 (2\omega^2 - a^2 - 2b + \kappa^2 + \phi^2)}{a^2 + (\phi + \kappa)^2 + 4\omega^2 + 2a(\phi + \kappa)\cos(\omega\Delta_i) - 4a\omega \sin(\omega\Delta_i)}. \tag{2.30}$$

Proof: When $\epsilon = 0$, this reduces back to the original solution of Equation (2.24), Δ_i. If we substitute the perturbation formulae (2.28) and (2.29) into Equation (2.19), we obtain

$$\begin{aligned}
0 &= -\lambda\xi\beta_t[\kappa + i\omega + \epsilon(ix + y)][\cos(\omega(\Delta_i + \epsilon\Delta_p)) - i\sin(\omega(\Delta_i + \epsilon\Delta_p))] \\
&\quad -N[\phi + i\omega + \epsilon(ix + y)][\kappa + i\omega + \epsilon(ix + y)] - \lambda\beta_c\eta \\
&= -\lambda\xi\beta_t[\kappa + \epsilon y - (\omega + \epsilon y)i][\cos(\omega(\Delta_i + \epsilon\Delta_p)) - i\sin(\omega(\Delta_i + \epsilon\Delta_p))] \\
&\quad -N[\phi + \epsilon y + (\epsilon x + \omega)i][\kappa + \epsilon y + (\epsilon x + \omega)i] - \lambda\beta_c\eta. \qquad (2.31)
\end{aligned}$$

By rearranging and putting the real and imaginary parts together, we have

$$\begin{aligned}
0 &= -\lambda\xi\beta_t[(\kappa + \epsilon y)\cos(\omega(\Delta_i + \epsilon\Delta_p)) + (\omega + \epsilon x)\sin(\omega(\Delta_i + \epsilon\Delta_p))] \\
&\quad -\lambda\beta_c\eta - N[(\phi + \epsilon y)(\kappa + \epsilon y) - (\epsilon x + \omega)^2] \\
&\quad -\lambda\xi\beta_t[(\omega + \epsilon x)\cos(\omega(\Delta_i + \epsilon\Delta_p)) - (\kappa + \epsilon y)\sin(\omega(\Delta_i + \epsilon\Delta_p))]i \\
&\quad -N(\phi + \kappa + 2\epsilon y)(\omega + \epsilon x)i. \qquad (2.32)
\end{aligned}$$

Now we view the RHS of (2.32) as a function of ϵ and use a Taylor expansion for small values of ϵ, then Equation (2.32) becomes

$$\begin{aligned}
&- \lambda\xi\beta_t[\kappa\cos(\omega\Delta_i) + \omega\sin(\omega\Delta_i)] - N(\phi\kappa - \omega^2) - \lambda\beta_c\eta \\
&- \lambda\xi\beta_t[(y + \omega^2\Delta_p)\cos(\omega\Delta_i) + (x - \kappa\omega\Delta_p)\sin(\omega\Delta_i)]\epsilon - N[(\phi + \kappa)y - 2\omega x)]\epsilon \\
&- \lambda\xi\beta_t[\omega\cos(\omega\Delta_i) - \kappa\sin(\omega\Delta_i)]i - N(\phi + \kappa)\omega i \\
&- \lambda\xi\beta_t[(x - \kappa\omega\Delta_p)\cos(\omega\Delta_i) - (y + \omega^2\Delta_p)\sin(\omega\Delta_i)]\epsilon i - N[(\phi + \kappa)x + 2\omega y)]\epsilon i \\
&+ O(\epsilon^2) \\
&= 0. \qquad (2.33)
\end{aligned}$$

Note that the first line and the third line of Equation (2.33) are both equal to 0 since they exactly match the first equation and the second equation in (2.20), respectively, and by definition of Δ_i, it satisfies (2.20). Therefore, Equation (2.33) reduces to

$$\begin{aligned}
& - && \lambda\xi\beta_t[(y+\omega^2\Delta_p)\cos(\omega\Delta_i)+(x-\kappa\omega\Delta_p)\sin(\omega\Delta_i)]\epsilon - N[(\phi+\kappa)y-2\omega x)]\epsilon \\
& - && \lambda\xi\beta_t[(x-\kappa\omega\Delta_p)\cos(\omega\Delta_i)-(y+\omega^2\Delta_p)\sin(\omega\Delta_i)]\epsilon i - N[(\phi+\kappa)x+2\omega y)]\epsilon i \\
& + && O(\epsilon^2) \\
& = && 0. \qquad (2.34)
\end{aligned}$$

The above Equation (2.34) holds for all small ϵ, implying that both the real and imaginary parts of the $O(\epsilon)$ terms are 0. This amounts to a system of linear equations in x and y

$$\begin{aligned}
0 &= a[(y+\omega^2\Delta_p)\cos(\omega\Delta_i)+(x-\kappa\omega\Delta_p)\sin(\omega\Delta_i)]+(\phi+\kappa)y-2\omega x \\
0 &= a[(x-\kappa\omega\Delta_p)\cos(\omega\Delta_i)-(y+\omega^2\Delta_p)\sin(\omega\Delta_i)]+(\phi+\kappa)x+2\omega y, \quad (2.35)
\end{aligned}$$

where $a = \lambda\xi\beta_t N^{-1}$ as defined before in (2.23).

Solving for x in terms of y by the first equation and substituting the resultant expression into the second equation in (2.35), we obtain

$$y = \frac{\Delta_p\omega^2(2\omega^2-a^2-2b+\kappa^2+\phi^2)}{a^2+(\phi+\kappa)^2+4\omega^2+2a(\phi+\kappa)\cos(\omega\Delta_i)-4a\omega\sin(\omega\Delta_i)}, \qquad (2.36)$$

as desired. ■

Based on Lemma 2.1, we have the following result that leads to the insights of why the smallest solution to (2.24) is enough for characterizing the stability condition.

Lemma 2.2 *The sign of the real part y is the same as the sign of the perturbation Δ_p.*

Proof: Consider the RHS of Equation (2.30), the denominator equals

$$\begin{aligned} &= a^2 + (\phi + \kappa)^2 + 4\omega^2 + 2a(\phi + \kappa)\cos(\omega\Delta_i) - 4a\omega\sin(\omega\Delta_i) \\ &= a^2 + (\phi + \kappa)^2 + 4\omega^2 + 2a\sqrt{(\phi + \kappa)^2 + (2\omega)^2}\cos(\omega\Delta - \alpha) \\ &\geq a^2 + (\phi + \kappa)^2 + 4\omega^2 - 2a\sqrt{(\phi + \kappa)^2 + (2\omega)^2} \\ &= \left(a - \sqrt{(\phi + \kappa)^2 + (2\omega)^2}\right)^2 \\ &\geq 0, \end{aligned} \tag{2.37}$$

where the second equality follows by setting

$$\alpha = \arctan\left(\frac{2\omega}{\phi + \kappa}\right);$$

and the numerator equals

$$\begin{aligned} &= \Delta_p\omega^2(2\omega^2 - a^2 - 2b + \kappa^2 + \phi^2) \\ &= \Delta_p\Omega(2\Omega + B) \\ &= \Delta_p\Omega\left(-B + \sqrt{B^2 - 4D} + B\right) \\ &= \Delta_p\Omega\sqrt{B^2 - 4D}, \end{aligned} \tag{2.38}$$

where the second equality follows from definition of B in (2.23), and the third equality is by the definition of Ω in (2.26).

Therefore, by (2.37) and (2.38) we can express the real part y of the perturbation as

$$y = \frac{\Delta_p\Omega\sqrt{B^2 - 4D}}{a^2 + (\phi + \kappa)^2 + 4\omega^2 + 2a\sqrt{(\phi + \kappa)^2 + (2\omega)^2}\cos(\omega\Delta - \alpha)}, \tag{2.39}$$

because of (2.37), $\Omega > 0$, and $\sqrt{B^2 - 4D} > 0$ (since $y \neq 0$), y has the same sign as Δ_p. ■

Lemma 2.2 means that for any solution Δ_i to the transcendental equation (2.24), the root r crosses the imaginary axis from left to right as the delay Δ increases beyond Δ_i. This implies that system instability that occurs after passing through the first Hopf curve remains as more Hopf curves are passed through. In other words, given all the other parameters, the instability remains when $\Delta \geq \Delta_1 = \Delta_{cr}$, not matter it is smaller or greater than any other solutions of (2.24), $\Delta_2, \Delta_3, ..., \Delta_\infty$ (there are an infinite number of solutions to (2.24)). So system stability only changes once at the point where $\Delta = \Delta_{cr}$.

Useful results based on Theorem 2.1

The key purpose of the paper is to study how disclosure of air pollution information affects the stability of traffic. To this end, we compared system performance with and without the air pollution information provided. Note that if there is no air pollution information provided, we can reasonably assume that motorists perceive no disutility of air pollution in their routing decision, that is: $\beta_c = 0$. This leads to the following straightforward Corollary of Theorem 2.1.

Corollary 2.1 *Suppose $\beta_c = 0$ in (2.7), if $\beta_t \leq \lambda^{-1}\xi^{-1}\phi N$, then equilibrium 2.8 is stable for all $\Delta > 0$; Otherwise, let*

$$\Delta_{cr} = \frac{1}{\sqrt{a^2 - \phi^2}} \arccos\left(-\frac{\phi}{a}\right), \tag{2.40}$$

then the equilibrium (2.8) is stable if $\Delta < \Delta_{cr}$ and unstable otherwise (where $a = \lambda\xi\beta_t N^{-1}$ is defined in (2.23)).

Proof: Note that $\beta_c = 0$ implies $b = 0$ by definition (2.17), so B and D defined in (2.23) become

$$B = \kappa^2 + \phi^2 - a^2; \; D = \phi^2\kappa^2 - a^2\kappa^2.$$

It follows that

$$B^2 - 4D \;\; = \;\; (\kappa^2 + \phi^2 - a^2)^2 - 4\phi^2\kappa^2 + 4a^2\kappa^2$$

$$
\begin{aligned}
&= (\phi^2 - \kappa^2 - a^2)^2 \\
&\geq 0. \qquad (2.41)
\end{aligned}
$$

Hence by Theorem 2.1, we have

$$
\begin{aligned}
\Omega^* &= \frac{-B + \sqrt{B^2 - 4D}}{2} \\
&= \frac{-\phi^2 - \kappa^2 + a^2 + |\phi^2 - \kappa^2 - a^2|}{2} \\
&= \begin{cases} -\kappa^2, & \text{if } \phi^2 \geq \kappa^2 + a^2, \\ a^2 - \phi^2, \text{otherwise}. \end{cases} \qquad (2.42)
\end{aligned}
$$

From (2.42) we know that the only case $\Omega > 0$ is when $a^2 > \phi^2$, i.e., $\beta_t > \lambda^{-1}\xi^{-1}\phi N$ by definition of a.

Therefore, by Theorem 2.1, if $\beta_t > \lambda^{-1}\xi^{-1}\phi N$, we can calculate Δ_{cr} as (2.40) which splits the whole domain of $\Delta \in [0, \infty)$ into two that characterize the stability of the equilibrium; Otherwise the equilibrium is always stable. ■

Now we show a second corollary of Theorem 2.1, which was applied to our numerical example. In that example, we approximated $g(Q_i)$ as an affine function that represents a simple (but essential) relationship between travel time and queue length; we approximated $h(Q_i)$ as an exponential function, which well captures the relationship between emission strength of $PM_{2.5}$ per vehicle on the link and the number of vehicles on the same link. So we have a result tailored to these functional forms $g(Q_i)$ and $h(Q_i)$.

Corollary 2.2 *Suppose $g(Q_i) = \alpha Q_i + \beta$ with $\alpha, \beta > 0$, and $h(Q_i) = \gamma \exp(\delta Q_i)$ with $\gamma > 0, \delta < 0$, then the equilibrium of system (2.7) is*

$$
Q_i = \frac{\beta\lambda}{N - \alpha\lambda};\ C_i = \frac{\beta\gamma\lambda}{\kappa(N - \alpha\lambda)} \exp\left(\frac{\beta\delta\lambda}{N - \alpha\lambda}\right) + C_b,\ i = 1, ..., N. \qquad (2.43)
$$

And in Theorem 2.1, the variables ϕ and η can be calculated as

$$\phi = \frac{(N - \alpha\lambda)^2}{\beta N^2}; \ \eta = \frac{\gamma N + \gamma(\beta\delta - \alpha)\lambda}{N - \alpha\lambda} \exp\left(\frac{\beta\delta\lambda}{N - \alpha\lambda}\right). \tag{2.44}$$

Proof: The equilibrium solution follows from the symmetry of (2.7):

$$\frac{\lambda}{N} - \frac{Q_i}{\alpha Q_i + \beta} = 0; \ \gamma Q_i \exp(\delta Q_i) - \kappa C_i = 0$$

$$\Rightarrow Q_i = \frac{\beta\lambda}{N - \alpha\lambda}; \ C_i = \frac{\beta\gamma\lambda}{\kappa(N - \alpha\lambda)} \exp\left(\frac{\beta\delta\lambda}{N - \alpha\lambda}\right).$$

Then the expressions in (2.44) can be obtained by substituting the equilibrium solution (2.43) into (2.14). ■

Further, we have an observation that offers insights on how the traffic demand (or system load) affects traffic stability. Recall that we had an observation from our numerical examples that the system stability characterization is invariant with respect to the ratio of λ/N, i.e., the total arrival rate divided by the number of alternatives (we call this quantity "normalized arrival rate"). This implies that when both λ and N change, they affect system stability only through their ratio λ/N. In fact, we can prove this invariance property rigorously.

Proposition 2.1 *Given a fixed normalized arrival rate λ/N, the stability condition for the equilibrium solution (2.8) is invariant.*

Proof: First note that quantity B and D are functions of a, b, κ and ϕ by (2.23), since κ does not depend on N nor λ, so B and D depend on N and λ only through a, b and ϕ. Then by Theorem 2.1, we know that system stability depends on N and λ only through a, b and ϕ.

By definitions (2.14) and (2.23), we have

$$\begin{aligned} a &= \beta_t \lambda N^{-1} g'(Q^*); \\ b &= \beta_c \lambda N^{-1} (h(Q^*) + h'(Q^*) \cdot Q^*); \\ \phi &= 1/g(Q^*) - g'(Q^*) Q^* / g(Q^*)^2. \end{aligned} \tag{2.45}$$

Then by symmetry of system (2.7), we know that the equilibrium solution Q^* is a root of the following function:

$$q(x) := \frac{x}{g(x)} - \frac{\lambda}{N}.$$

Since $g(\cdot)$ is an exogenous function independent of N and λ, it follows that Q^* depends on N and λ only via the ratio λ/N, so do $g(Q^*)$ and $g'(Q^*)$. Similarly, we have $h(Q^*)$ and $h'(Q^*)$ depend on N and λ only through λ/N. Therefore, we deduce by (2.45) that all three quantities a, b and ϕ depend on N and λ only through λ/N. This gives us the result. ■

Finally, we prove a sufficient condition on the intuitive result that the more the users value travel time, the less stable the system becomes. We observed this pattern in the numerical example. To simplify the statement of the following proposition and also Theorem 2.2 in next section, we let $\Delta_{cr} = \infty$ if the system is stable for all $\Delta > 0$.

Proposition 2.2 *If the travel time function is strictly increasing in queue length at the equilibrium solution Q^* defined in (2.8), i.e.*

$$\frac{d(g(x))}{dx}\Big|_{x=Q^*} > 0, \tag{2.46}$$

then Δ_{cr} is strictly decreasing in $\beta_t > 0$ if $\Delta_{cr} < \infty$.

Proof: By definition of ξ in (2.14), we know that condition (2.46) is equivalent to $\xi > 0$. Suppose $\Delta_{cr} < \infty$ (i.e., $\Omega^* \in \mathbb{R}_{++}$ by Theorem 2.1), then by (2.26) and $\phi > 0$ (by the definition of Q^* in (2.8)), we have

$$\begin{aligned}
\sqrt{B^2 - 4D} &= \sqrt{(\phi^2 - \kappa^2 - a^2)^2 - 4b(\phi^2 - \kappa^2 - a^2) - 8b\phi\kappa} \\
&\leq \sqrt{(\phi^2 - \kappa^2 - a^2)^2 - 4b(\phi^2 - \kappa^2 - a^2)} \\
&\leq \sqrt{(\phi^2 - \kappa^2 - a^2)^2 - 4b(\phi^2 - \kappa^2 - a^2) + 4b^2} \\
&= \sqrt{(\phi^2 - \kappa^2 - a^2 - 2b)^2} \\
&= |\phi^2 - \kappa^2 - a^2 - 2b|. \qquad (2.47)
\end{aligned}$$

Then we have by (2.26) that

$$\begin{aligned}
\Omega^* &= \frac{-B + \sqrt{B^2 - 4D}}{2} \\
&\leq \frac{-\phi^2 - \kappa^2 + a^2 + 2b + |\phi^2 - \kappa^2 - a^2 - 2b|}{2} \\
&= \begin{cases} -\kappa^2, & \text{if } 2b \leq \phi^2 - \kappa^2 - a^2, \\ a^2 + 2b - \phi^2, & \text{otherwise.} \end{cases} \qquad (2.48)
\end{aligned}$$

Therefore, in order to have $\Omega^* \in \mathbb{R}_{++}$, it must be true that

$$2b > \phi^2 - \kappa^2 - a^2. \qquad (2.49)$$

Note that by (2.27) in Theorem 2.1, and since $a = \lambda\beta_t\xi N^{-1}$ by (2.23) and $\xi > 0$, we have Δ_{cr} strictly decreasing in β_t if we can show Ω^* strictly increasing in a. By (2.26) the partial derivative of Ω^* with respect to a is

$$\begin{aligned}
\frac{\partial\Omega^*}{\partial a} &= \frac{1}{2}\frac{\partial\left(-B + \sqrt{B^2 - 4D}\right)}{\partial a} \\
&= \frac{1}{2}\frac{\partial\left(-\phi^2 - \kappa^2 + a^2 + 2b + \sqrt{(\phi^2 + \kappa^2 - a^2 - 2b)^2 - 4(\phi\kappa + b)^2 + 4a^2\kappa^2}\right)}{\partial a} \\
&= a + \frac{1}{2}\frac{\partial\left(\sqrt{(\phi^2 + \kappa^2 - a^2 - 2b)^2 - 4(\phi\kappa + b)^2 + 4a^2\kappa^2}\right)}{\partial a}
\end{aligned}$$

$$
\begin{aligned}
&= a + \frac{1}{4}\frac{-4a(\phi^2+\kappa^2-a^2-2b)+8a\kappa^2}{\sqrt{(\phi^2+\kappa^2-a^2-2b)^2-4(\phi\kappa+b)^2+4a^2\kappa^2}} \\
&= a + \frac{-a(\phi^2-\kappa^2-a^2-2b)}{\sqrt{(\phi^2+\kappa^2-a^2-2b)^2-4(\phi\kappa+b)^2+4a^2\kappa^2}} \\
&= a\left(1+\frac{2b-\phi^2+\kappa^2+a^2}{\sqrt{(\phi^2+\kappa^2-a^2-2b)^2-4(\phi\kappa+b)^2+4a^2\kappa^2}}\right) \\
&> a, \qquad (2.50)
\end{aligned}
$$

where the inequality follows by (2.49). Therefore, $\partial\Omega^*/\partial a > 0$ and we have the result. ■

Note that in the traffic flow setting, it is true that the higher the vehicle density, the lower the traffic speed. Hence the travel time function $g(Q_i)$ is indeed strictly decreasing in the queue length Q_i. Hence the result in Proposition 2.2 is general in our traffic network analysis.

2.6.3 Proof of Theorem 2.2

Proof: First note that condition (2.10) is equivalent to $\eta > 0$ by the definition of η in (2.14).

Proof for the first part of the conclusion is similar to the proof of Proposition 2.2. Suppose $\Delta_{cr} < \infty$ (i.e., $\Omega^* \in \mathbb{R}_{++}$ by Theorem 2.1), then by (2.26) the term inside the square root $\sqrt{B^2-4D}$ must be nonnegative in the first place, i.e.,

$$
\begin{aligned}
B^2-4D &= (\phi^2+\kappa^2-a^2-2b)^2-4(\phi\kappa+b)^2+4a^2\kappa^2 \\
&= (\phi^2-\kappa^2-a^2)^2-4b(\phi^2-\kappa^2-a^2)-8b\phi\kappa \\
&\geq 0. \qquad (2.51)
\end{aligned}
$$

Note that by (2.27) in Theorem 2.1, plus $b = \lambda\beta_c\eta N^{-1}$ by (2.23) and $\eta > 0$, if we can show that Ω^* is decreasing in b then we have Δ_{cr} strictly increasing in β_c. By (2.26) the partial derivative of Ω^* with respect to b is

$$\begin{aligned}
\frac{\partial \Omega^*}{\partial b} &= \frac{1}{2}\frac{\partial\left(-B + \sqrt{B^2 - 4D}\right)}{\partial b} \\
&= \frac{1}{2}\frac{\partial\left(-\phi^2 - \kappa^2 + a^2 + 2b + \sqrt{(\phi^2 + \kappa^2 - a^2 - 2b)^2 - 4(\phi\kappa + b)^2 + 4a^2\kappa^2}\right)}{\partial b} \\
&= 1 + \frac{1}{2}\frac{\partial\left(\sqrt{(\phi^2 + \kappa^2 - a^2 - 2b)^2 - 4(\phi\kappa + b)^2 + 4a^2\kappa^2}\right)}{\partial b} \\
&= 1 + \frac{1}{2}\frac{\partial\left(\sqrt{(\phi^2 - \kappa^2 - a^2)^2 - 4b(\phi^2 - \kappa^2 - a^2) - 8b\phi\kappa}\right)}{\partial b} \\
&= 1 - \frac{(\phi^2 - \kappa^2 - a^2) + 2\phi\kappa}{\sqrt{(\phi^2 - \kappa^2 - a^2)^2 - 4b(\phi^2 - \kappa^2 - a^2) - 8b\phi\kappa}}. \qquad (2.52)
\end{aligned}$$

Hence showing $\partial\Omega/\partial b < 0$ is equivalent to showing

$$4\phi\kappa(\phi^2 - \kappa^2 - a^2) + 4\phi^2\kappa^2 + 4b(\phi^2 - \kappa^2 - a^2) + 8b\phi\kappa > 0. \qquad (2.53)$$

We divide the LHS of (2.53) by a factor of 4 and obtain

$$\begin{aligned}
& \phi\kappa(\phi^2 - \kappa^2 - a^2) + \phi^2\kappa^2 + b(\phi^2 - \kappa^2 - a^2) + 2b\phi\kappa \\
= \; & (\phi\kappa + b)(\phi^2 - \kappa^2 - a^2) + \phi^2\kappa^2 + 2b\phi\kappa \\
> \; & \left(\phi\kappa + \frac{\phi^2 - \kappa^2 - a^2}{2}\right)(\phi^2 - \kappa^2 - a^2) + \phi^2\kappa^2 + (\phi^2 - \kappa^2 - a^2)\phi\kappa \\
= \; & 2\phi\kappa(\phi^2 - \kappa^2 - a^2) + \phi^2\kappa^2 + \frac{1}{2}(\phi^2 - \kappa^2 - a^2)^2 \\
\geq \; & 2\phi\kappa(\phi^2 - \kappa^2 - a^2) + \phi^2\kappa^2 + 2b(\phi^2 - \kappa^2 - a^2) + 4b\phi\kappa \\
> \; & 2\phi\kappa(\phi^2 - \kappa^2 - a^2) + \phi^2\kappa^2 + (\phi^2 - \kappa^2 - a^2)^2 + 4b\phi\kappa \\
= \; & (\phi\kappa + \phi^2 - \kappa^2 - a^2)^2 + 4b\phi\kappa \\
\geq \; & 0, \qquad (2.54)
\end{aligned}$$

where the first and the third inequalities follow by (2.49) since $\Omega \in \mathbb{R}_{++}$ and $\phi > 0$; the second inequality follows by (2.51); and the last inequality is due to $\phi > 0$.

Therefore, Δ_{cr} is strictly increasing in β_c if Δ_{cr} is finite. This completes the proof of the first part of the conclusion.

Now we show the second part of the conclusion by contradiction. Suppose $\Delta_{cr}^{\text{no}} = \infty$ but $\Delta_{cr}^{\text{yes}} < \infty$. Recall that when $b = 0$, (2.42) in Corollary 2.1 implies that $\Omega^* \notin \mathbb{R}_{++}$ if only if $\phi^2 \geq a^2$, i.e., $\Delta_{cr}^{0} = \infty$ if only if $\phi^2 \geq a^2$. Now it follows from $\phi^2 \geq a^2$ and $\phi, \kappa > 0$ that $\phi\kappa + b > a\kappa$ when $b > 0$, which implies

$$\begin{aligned}
\Omega^* &= -B + \sqrt{B^2 - 4D} \\
&= -\phi^2 - \kappa^2 + a^2 + 2b + \sqrt{(\phi^2 + \kappa^2 - a^2 - 2b)^2 - 4(\phi\kappa + b)^2 + 4a^2\kappa^2} \\
&< -\phi^2 - \kappa^2 + a^2 + 2b + \sqrt{(\phi^2 + \kappa^2 - a^2 - 2b)^2} \\
&= -\phi^2 - \kappa^2 + a^2 + 2b + |\phi^2 + \kappa^2 - a^2 - 2b|. \qquad (2.55)
\end{aligned}$$

Since we assume $\Delta_{cr}^{\text{yes}} < \infty$, then the above inequality implies

$$2b > \phi^2 + \kappa^2 - a^2 > 0, \qquad (2.56)$$

where the second inequality is due to $\phi^2 \geq a^2$ and $\kappa > 0$.

Finally, we have

$$\begin{aligned}
B^2 - 4D &= (\phi^2 + \kappa^2 - a^2 - 2b)^2 - 4(\phi\kappa + b)^2 + 4a^2\kappa^2 \\
&< (2b)^2 - 4(\phi\kappa + b)^2 + 4a^2\kappa^2 \\
&= -4\phi^2\kappa^2 + 4a^2\kappa^2 - 8b\phi\kappa \\
&\leq 0, \qquad (2.57)
\end{aligned}$$

where the first inequality follows by (2.56) and the last inequality is due to $\phi^2 \geq a^2$ and $\phi \geq 0$.

Therefore, (2.57) is a contradiction to $\Omega^* \in \mathbb{R}$, hence a contradiction to Δ_{cr}^{yes} being finite. This completes the proof. ■

2.7 Appendix B: Details of Numerical and Simulation Analysis

2.7.1 Data for Numerical Analysis

In this section, we provide the supporting data for the numerical analysis in the result section of the main paper. The example considers two separate parallel one-mile long single-lane links, so the numerical value of average traffic density ρ (veh/mile) equals that of the number of vehicles on the lane, Q (veh). Table 2.3 shows all the data we used for our numerical analysis. The sensitivity analysis in the main paper was done based on these baseline values listed in Table 2.3.

Parameter	**Symbol**	**Value**
Number of links (single lane)	N	2
Length of links	L	1 mile
Total arrival rate	λ	30 veh/min
Free-flow speed	v_{free}	60 mph
Cross-sectional area of the "air box"	A	15 m^2
Dilution rate constant	κ	6 /min
$\text{PM}_{2.5}$ background concentration	C_b	35 $\mu\text{g/m}^3$
WTP of travel time	β_t	0.6 \$/min
WTP of $\text{PM}_{2.5}$ concentration	β_c	0.2 \$/(μg/m^3)
Travel time function	$g(Q)$	$0.0526Q + 1.0$ (min)
$\text{PM}_{2.5}$ missions strength function	$h(Q)$	$1.176e^{-0.0058Q}$ (μg/m^3/s/veh)

Table 2.3: Data for numerical analysis

In the above table, N, L are assumed data. A is obtained by assuming that the longitudinal "box" surrounding the lane has a cross section of a 5-meter wide and 3-meter tall rectangle (which is sufficient to cover the whole lane and the vehicles on it). κ^p is estimated based on wind speed: assuming that horizontal wind ventilation dominates the vertical one and has an average speed of 0.5 m/s with direction normal to the traffic direction, so every minute the air inside the "box" is exchanged for 60/(5/0.5) = 6 times. The background concentration of $\mathrm{PM}_{2.5}$ is assumed to be the U.S. National Ambient Air Quality Standards of $35\,\mu\mathrm{g/m}^3$ for 24-hour average $\mathrm{PM}_{2.5}$ concentration. Willingness to pay for travel time is estimated as 36 \$/hr for the commuters which is equivalent to 0.6\$/min. We assume that commuters' willingness to pay for reduction in $\mathrm{PM}_{2.5}$ concentration is one third of the WTP for travel time (min) savings, so $\beta_c = 0.2$ \$/(μg/m^3) . The two functions of queue length $g(Q)$ and $h(Q)$ are fitted based on analytic models in the literature.

Specifically, we use a linear fit for the data points (travel time against traffic density) to obtain the average travel time function $g(Q)$. The data points shown in Figure 2.12(a) are generated based on the classic Greenberg model [54] that describes the relationship between average traffic speed v and traffic density ρ (we divide the link length by traffic speed to obtain the travel time). This model includes an explicit form of the relationship between v and ρ as

$$\rho(v) = \rho_j \exp(-v/v^*) \Leftrightarrow v = v^* \ln(\rho_j/\rho) \tag{2.58}$$

where the optimal average speed v^* is 16.1 mph; and the jam density ρ_j is 215 veh/mile for a single lane measured at a typical road segment. Then we evaluate formula (2.58) as density observations (here the density has the same numerical value as that of the number of vehicles on the 1-mile long lane, Q) together with calculated $T = 60/v$ (min) at discrete average speed points every 5 mph from 5 mph to 60 mph. For simplicity, we fit an affine function $g(Q) = \alpha Q + \beta$ to these twelve observations (Q, T) to approximate the travel time function $T \approx g(Q)$. We restrict the intercept $g(0) = 1.0$ to represent free-flow travel time of 1 min at a free-flow speed of 60 mph. From (2.58) we can see that as traffic density approaches 0, traffic speed according to the Greenberg model approaches infinity, which is unrealistic. Hence the restriction of $g(0) = 1.0$ is useful also because it helps overcome this limitation of the Greenberg model.

Figure 2.12(a) shows the fitted affine function $g(Q) = 0.0526Q + 1.0$ on the data points (Q, T), which achieves a satisfactory goodness of fit ($R^2 = 0.87$).

For the emission strength function, we fit a function to the data points generated based on the density-speed relationship (2.58) as well as an analytic form that approximates the emission rate (as a function of fleet speed) data from the MOtor Vehicle Emission Simulator (MOVES) software developed by the EPA for a realistic fleet composition [19]. The original emission rate function to be fitted takes the following form [19]

$$r(v) = \exp\left(\sum_{j=1}^{5} c_j v^{(j-1)}\right), \tag{2.59}$$

where $c_1, ..., c_5$ are pollutant specific model coefficients which take values: $c_1 = -1.223$, $c_2 = -0.1769$, $c_3 = 0.00664$, $c_4 = -0.00011$, $c_5 = 6.724 \times 10^{-7}$ for $\mathrm{PM}_{2.5}$ [19]; and $v^{(j-1)}$ denotes the $(j-1)^{\text{th}}$ power of v. Based on (2.59) and the expression in (2.58), we use similar approach to fit the function $h(Q) = A^{-1}L^{-1}r(v(Q))v(Q)$ (μg/m^3/s). We evaluate (2.58) for $Q = \rho(v)$ and the corresponding expression $H = A^{-1}L^{-1}r(v(Q))v(Q)$ at discrete speed points every 5 mph from 5 mph to 60 mph. Then we get twelve data points of (Q, H) for which we fit the function $H \approx h(Q)$. Now since the function $h(Q)$ for $\mathrm{PM}_{2.5}$ turns out to be decreasing in Q, so instead of a linear function, we choose to approximate it with an exponential function $h(Q) = \gamma \exp(\delta Q)$ with $\gamma > 0, \delta < 0$ so as to ensure it is always positive. Note that the emission strength increases first as speed increases and then starts to decrease slightly when speed increases to about 50 mph. This cannot be represented by our simple exponential fit since the exponential function is monotone and has sharper decrease for smaller Q (where speed is high). Hence we compromise for this by restricting the intercept of the exponential function to be equal to the emission strength at the free-flow speed (60 mph), which is $h(0) = 52.9$. Figure 2.12(b) shows the fitted exponential function $h(Q) = 52.9 \exp(-0.0058Q)$ on the data points (Q, H), which achieves a satisfactory goodness of fit ($R^2 = 0.96$). An important note is that although the fitted $h(Q)$ is monotonically decreasing in Q, the total emission source strength $S(Q) = h(Q) \cdot Q$ is monotonically increasing in Q within normal average speed range (5 mph ~ 60 mph, corresponds to about Q from 0 to 160), as shown in Figure 2.13.

With the functional forms fitted for $g(Q)$ and $h(Q)$ in our numerical studies, we can use Corollary 2 to verify the equilibrium solution of system (2.7) and examine its convergence.

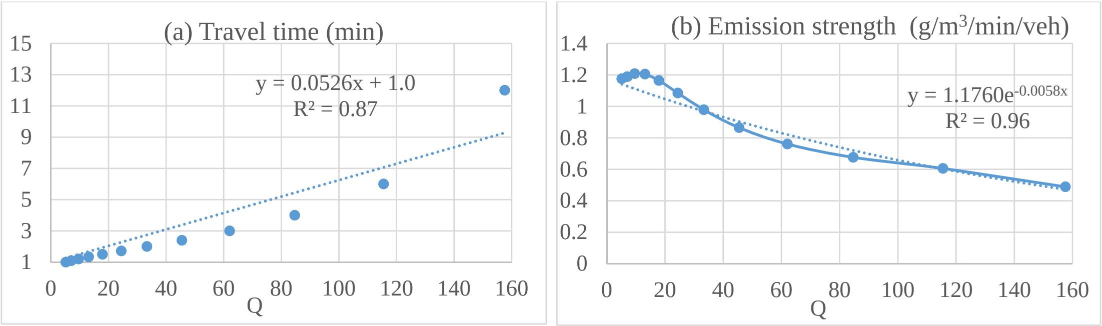


Figure 2.12: Approximating travel time and emission strength functions $g(Q)$ and $h(Q)$ (dots: data points; dashed curves: fitted functions.

2.7.2 Simulation Model, Data and Additional Discussion

In the main paper, we demonstrate the traffic stability enhancing effect of air pollution information disclosure via simulation of traffic on the GWB. In this section, we describe the simulation model and data that we used for our traffic

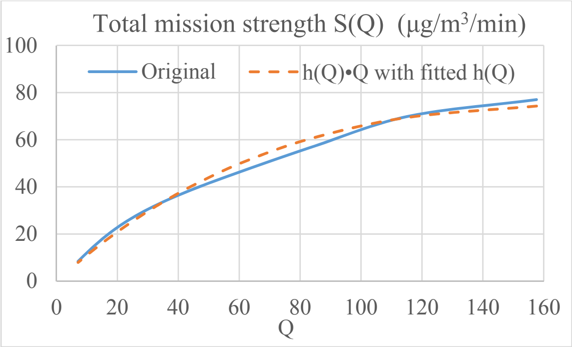


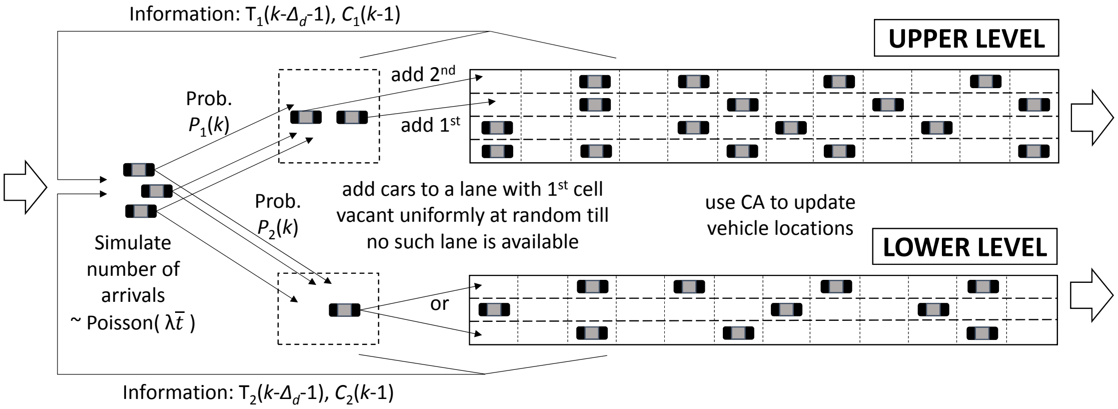
Figure 2.13: Total emission source strength function $S(Q)$ (original and fitted).

simulation of the GWB. Moreover, additional simulation results and discussion are also provided.

Simulation Model

For micro-simulation of traffic, we use a popular stochastic traffic simulation model – Cellular Automaton (CA) [92], which is well known to be able to reproduce the key features of traffic flow such as shock waves and different phases in traffic fundamental diagram that are observed in reality [92]. Despite many variants and extensions of the CA model, in this study, we adopt the original CA model [92] for an open road segment. However we embed it into the multi-link dynamic choice model, which is the key setting of this study. In particular, the GWB has one upper level and one lower level available for the traffic, so we model it as a two-link system. To use the CA [92] model, we define a one-dimensional array of M sites with open boundary conditions for each lane of each link. Each site may either be occupied by exactly one vehicle or unoccupied. Each vehicle has an integer valued velocity between zero and $v_{\max}$. We conduct discrete time simulation over a predefined total number of time steps

K. Each time step represents a length of $\bar{t}$ (s). At every simulation time step $k = 1, ..., K$, we simulate the vehicle arrivals, route choices, vehicle movements as well as vehicle departures if any. Figure 2.14 shows the whole process of the simulation at every step k.



Information: $T_1(k-\Delta_d-1)$, $C_1(k-1)$
UPPER LEVEL
add 2nd
add 1st
Prob. $P_1(k)$
Prob. $P_2(k)$
add cars to a lane with 1st cell vacant uniformly at random till no such lane is available
use CA to update vehicle locations
LOWER LEVEL
Simulate number of arrivals ~ Poisson($\lambda\bar{t}$)
or
Information: $T_2(k-\Delta_d-1)$, $C_2(k-1)$

Figure 2.14: Simulation process at each time step k.

Specifically, at each step k, we do the following. We first generate a random variable $N_{arr}(k) \sim \text{Poisson}(\lambda\bar{t})$ which represents the total number of new arrivals at step k, i.e., during time interval $[k\bar{t}, (k+1)\bar{t})$. Here rate parameter λ stands for the expected number of arrivals during a unit time interval (1s). Suppose the choice probability for link i is p_i, then each newly arrived vehicle is assigned to link i with probability p_i. These assigned vehicles are first put into the buffers (each link has an associated buffer with some sufficient length L_b) and then loaded onto the lanes of the links in a FIFO (first in first out) manner. The vehicles in the buffer of each link i (if any) are added sequentially, each to a lane whose first cells are vacant uniformly at random until no such lane is available for this link. The location updating of the vehicles on each link $i = 1, ..., N$ follows the procedure described below (where $v(s)$ and $loc(s)$ denote the speed of vehicle s and site index where vehicle s is located on the lane).

(1) For every newly added vehicle s,

$$v(s) \leftarrow \min[loc(1^{\text{st}} \text{ vehicle on the lane}) - 1,\ v_{\max}].$$

(2) For the last vehicle s on each lane, $v(s) \leftarrow \min[v(s) + 1,\ v_{\max}]$.

(3) For each of all other vehicles s,

$$v(s) = \min[v + 1,\ loc(\text{the vehicle ahead of } s) - loc(s) - 1,\ v_{\max}].$$

(4) For every vehicle s, $v(s) \leftarrow \max[v(s) - 1,\ 0]$ with probability p_d.

(5) For each vehicle s, $loc(s) \leftarrow loc(s) + v(s)$, and s leaves the lane if $loc(s) > L$.

The operation of taking the minimum in steps (1)~(3) essentially encodes how the drivers decide their speed: each driver chooses to accelerate by one unit speed level as long as the maximum speed is not surpassed and the vehicle remains behind the vehicle ahead to prevent collision. Step (4) applies random deceleration independently to all the vehicles according to a common deceleration probability p_d. This random slow-down is crucial in simulating realistic traffic flow since otherwise the dynamics is completely deterministic. It is used to model natural speed fluctuations as a result of human behavior or varying external conditions [92]. Note that we assume the vehicle that is about to enter the lane already accelerates to at least $v_{\max} - 1$, so it takes the speed as specified in step (1) when it enters the lane. We also assume that free-flow condition prevails downstream the link outlet, so the last vehicle on the lane will choose to accelerate or maintain maximum speed except for random slow-downs due to disturbances. In addition, since we focus on rush-hour traffic scenario for the GWB that has relatively strict traffic regulations such as low speed limit (45 mph) [6], we do not simulate the lane changing behavior.

Notice that other than the random slow-down due to unknown disturbances (proposed in original CA model), the stochasticity in our traffic simulation also includes the following aspects: 1) number of arrivals is random; 2) link choices are random; 3) the vehicles in the buffer are added to different lanes randomly.

As shown in Figure 2.14, a key component of our simulation experiment is information provision that affects the link choice behavior. We simulate this by calculating the travel time and pollutant concentration and provide the result as feedback information to the system. Link choice probabilities p_i are then updated every simulation step based on the Multinomial Logit (MNL) model. Specifically, the following procedure is adopted. Note that in our discrete time simulation framework, the information at least "lagged" for one time interval, since the best we can do to simulate the "real-time" information is to evaluate the output information for the previous time interval and distribute this information to users that arrive during the current time interval.

(1) The travel time is calculated based on the average speed of the vehicles on each link, which is usually the working principle of either fixed or mobile traffic sensors. To represent possible lag in travel time information, let $\Delta_d > 1$ be the number of discrete simulation intervals that equals $\Delta/\bar{t}$ (suppose Δ is an integer multiple of $\bar{t}$). So the average travel time evaluated for time interval k to be posted to inform new arrivals during time interval $k + 1$ is

$$\hat{T}_i(k) = T_i(k - \Delta_d) = \frac{M\bar{l}}{\text{average vehicle speed of link } i \text{ at step } (k - \Delta_d)}, \tag{2.60}$$

where the average vehicle speed on link i is calculated as the average speed of all the vehicles on link i and the vehicles waiting in the buffer of link i. Since each vehicle s waiting in the buffer has discretized speed

$v(s) = 0$), the average speed of link i is the average speed of the vehicles actually traveling on the link times the number of vehicles traveling on the link divided by the total number of vehicles both traveling on the link and waiting in the buffer of the link.

(2) Air pollutant concentration is modeled by discretizing the ordinary differential equation (ODE) in (2.7) plus an additive error term as below (for each $i = 1, ..., N$)

$$\begin{aligned} C_i(k+1) &= C_i(k) + \bar{t}[S_i(k) - \kappa(C_i(k) - C_b)] + \varepsilon(k) \\ &= \bar{t}S_i(k) + (1 - \bar{t}\kappa)C_i(k) + \bar{t}\kappa C_b + \varepsilon_i(k), \end{aligned} \tag{2.61}$$

where $S_i(k)$ is the total emission source strength on link i at time step k, which is estimated based on the average vehicle on link i and the emission rate function in the form of (2.59) at time k. $\varepsilon_i(k) \sim N(0, \sigma^2)$ represents any modeling error and measurement noise, we assume it is independent and identically distributed across different time intervals and links. And the reciprocal of the dilution rate constant κ has the same unit as $\bar{t}$. Note that in order to have stable system (2.61) we need to make sure that the simulation time scale is properly determined such that $\bar{t}\kappa < 1$.

(3) The link choice probabilities of each arriving vehicle are predicted by the MNL formula and updated for each time step as

$$p_i(k+1) = \frac{\exp(-\beta_t \hat{T}_i(k) - \beta_c C_i(k))}{\sum_{j=1}^{N} \exp(-\beta_t \hat{T}_j(k) - \beta_c C_j(k))}. \tag{2.62}$$

With these calculations, we construct a closed-loop structure for traffic simulation in which the arrivals to each link in the current time step affects future arrivals to the links due to the travel cost information (e.g., travel time, air pollution) feedback to the new users.

Simulation Data

To conduct the simulation for the GWB east bound morning traffic (from New Jersey to New York), we need relevant data for determining the simulation parameters. The data we used is given in Table 2.4. Moreover, the reference that directly supplies a data value is indicated in the bracket beside that value.

Variable	**Value**
Number of links (single lane), N	2
Number of lanes	upper level: 4; lower level: 3 [6]
Poisson rate of total arrivals, λ	2.477
Maximum speed in simulation, $v_{\max}$	3
Length of each simulation step $\bar{t}$	1 s
Number of sites on each lane M	193
Length of each lane site $\bar{l}$	7.5 m
Length of the buffer of each link L_b	60 m
Cross-sectional area of the "air box", A	60 m^2
Random slow-down probability, p_d	0.3
Dilution rate constant, κ	0.164/s
$PM_{2.5}$ background concentration, C_b	11.0 $\mu g/m^3$ [96]
Standard deviation of white noise, σ^2	1.0 $\mu g/m^3$
WTP of travel time, β_t	1.4 \$/min [18]
WTP of $PM_{2.5}$ concentration, β_c	0.14 \$/($\mu g/m^3$)
Emission rate function of $PM_{2.5}$	the same as (2.59) [19]
Delayed steps of travel time information, Δ_d	300

Table 2.4: Data used for the simulation of the GWB

In Table 2.4, the number of sites of each lane is calculated by dividing the lane length (1450 m [6]) by the length of each site (7.5m) [92], which yields $M = 193$. We assume that the length of the buffer is 60 m for each link (which is enough to accommodate the waiting vehicles in front of the link given the traffic demand in our simulation). Note that we suppose that the free-flow speed of the vehicles is 50 mph, which is slightly higher than the speed limit of 45 mph on the GWB. The maximum speed and simulation time scale can be determined one after the other. Here we first set the simulation step length as $\bar{t} = 1$ sec, then the maximum speed in the simulation is calculated by converting the 50 mph free-flow speed into the closest integer number of sites per simulation step (1 sec), which yields $v_{\max} = 3$. The expected number of total arrivals is calculated based on the hourly (7:00 AM to 8:00 AM) automobile volume data (8917 veh/hr) on the east bound of the GWB [2], which yields $\lambda = 2.477$. In our simulation, we assume the vehicles are all automobiles (the fractions of buses and trucks are less than eight percent [2]).

The cross-sectional area of the long "air box" covering the lanes ($A = 60\text{m}^2$) is obtained by assuming it as a 15-meter wide and 4-meter tall rectangle cross section (based on the clearance of the bridge and the width of motor vehicle lanes [6]). Using similar approach as we estimated κ in the numerical analysis, we suppose horizontal convection dominants the dispersion of $PM_{2.5}$. So the dilution rate constant for $PM_{2.5}$ is estimated mainly on the basis of the average wind speed of 7.8 mph (about 3.486m/s) on March 2017 [5], assuming the angle of the wind to the GWB is 45° on average. This gives us an estimate of $\kappa = (15\sqrt{2}/3.486)^{-1} = 0.164/s$. The $PM_{2.5}$ background concentration is set as the monthly average of 7:00 AM –8:00 AM measurement records at station IS-143 near the east end of the GWB throughout March 2017 [96].

The WTP for saving one minute commuting time is $\beta_t = 1.4\$/\text{min}$, according to the estimation by a recent study for New Yorkers [18]. However, due to lack of data, we assume the WTP for one $\mu\text{g}/\text{m}^3$ $PM_{2.5}$ on-road concentration reduction is only about one tenth of the WTP for one minute travel time saving, so $\beta_c = 0.1\beta_t = 0.14\ \$/\mu\text{g}/\text{m}^3$. Note that $\beta_c = 0.14\$/(\mu\text{g}/\text{m}^3)$ is similar to the estimate by a study in China [137] if we assume commuters evaluate their daily exposure based on the perceived concentration of their two commuting trips. Since β_c is relatively uncertain, we provide results under different values of β_c in next subsection.

Now we describe the initial conditions used in the simulation. At time step $k = 0$, there are 25 vehicles located randomly at each lane of both links. There is no vehicles waiting in the buffers. All the vehicles are moving very slowly and have speed $v = 0$ (since only three discrete levels of speed is used in the simulation). The initial $PM_{2.5}$ concentration is calculated as the equilibrium solution (see (2.43)) assuming a uniform traffic speed of 1 mph. The initial travel time information for both lanes is M (s), which is used upto simulation step $k = \Delta_d$. When $k > \Delta_d$, the truly evaluated information $\hat{T}_i(k-1) = T_i(k - \Delta_d - 1)$ will be posted. We start collecting data for simulation results evaluation at $k = \Delta_d$ and continue for 2400 steps (i.e., a 40 min-long period).

Figure 2.15 shows the simulated vehicle trajectories on one lane over the time steps 300 ∼ 500 in one sample path. This is a commonly used "time-space" plot in traffic engineering. The horizontal axis is the number of sites that indicates the distance from the link entrance; and the vertical axis is the time. Each small hollow black square at point (x, t) represents a vehicle at the corresponding site $x \in \{1, ..., M\}$ at time t. Each line running from bottom left to top right formed by connecting black hollow squares depicts the "time-space" trajectory of a vehicle on the link. The gap between two adjacent lines represents the headway between two adjacent vehicles.

From Figure 2.15, we observe that most of the vehicle trajectories are straight lines running from bottom left to top right, whose slopes are quite close to the maximum vehicle speed ($v_{\max} = 3$) specified in our simulation. This indicates that the vehicles are moving smoothly towards the end of the lane during most of their trips. But there are dark regions due to closely located black squares, indicating very slow traffic (even break down) within certain time windows. They correspond to temporary traffic jams caused by a stopped vehicle (due to random slow-down) in the front and the subsequent slowing down and stop of the cars following it. This represents the stop-and-go traffic that generates the so called "shock waves" along the lane. For example, beginning at time point 10, a vehicle stopped at about site 70, then the vehicles following it stopped one after another, causing a long traffic jam that lasted for about 40s and extends backwards to site 40. Notice that each shock wave (the dark region in Figure 2.15) moves backwards (from link exit to entrance) as time goes on. This is consistent with traffic flow theory and real observations in practice [92]. These realistic observations therefore also helps validate our simulation model. The speed at which these shock waves move upstream can also be used to calibrate our simulation model (such as parameters $v_{\max}$ and p) if real measured trajectory data is available.

We also note that the lines of vehicle trajectories are not evenly separated, they are denser some time and looser some other times. This reflects the varying arrival rate to this lane over time due to the fact that vehicles are assigned dynamically to each link and each lane based on the time-varying choice probabilities. This observation again adds reality-check to the functionality of our simulation model. Intuitively, denser lines (in the beginning part of the lanes) for a certain link are likely due to higher choice probability for that link, e.g., those peaks of the choice probability curves in Figure 2.6.

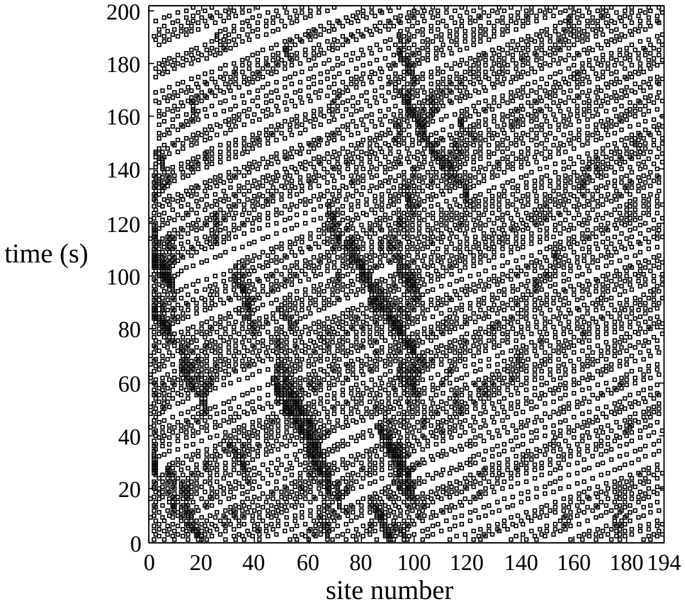


Figure 2.15: vehicle locations on one lane over a 200s time period in one sample path of the simulation.

Additional Discussion

The main paper includes comparisons of the simulation results between cases with and without air pollution information under the parameters fixed at the base values shown in Table 2.4. In the following, we also expand the comparisons to scenarios of different travel time information delays (Δ_d) and the WTP parameters (β_c), the two key types of parameters in this study.

The first piece of additional experiment is to examine the traffic performance by shortening or prolonging the lag Δ_d (as defined in Table 2.4) to mimic possible improvement or loss in the timeliness in the provision of travel time information. Figures 2.16 and 2.17 show the resultant number of vehicles and pollutant concentrations on two links under $\Delta_d = 180$ ($\Delta = 3$ min) and $\Delta_d = 480$ ($\Delta = 8$ min), respectively. Figure 2.18 compares the the total number of vehicles and the number of vehicles traveling on the lanes under the latter scenario. We can see that with shorter delay $\Delta_d = 180$ in travel time information, on average the traffic densities and $PM_{2.5}$ concentrations can gradually converge to stable conditions even without $PM_{2.5}$ concentration information posted. In this case posting pollution level information makes the system converging to equilibrium much faster. However, when timeliness in travel time information reduces significantly with Δ_d increased to 480, system performance without posting of $PM_{2.5}$ pollution information deteriorates dramatically. We can see from the lower two plots in Figure 2.17 that the magnitude of the traffic density oscillations become bigger and bigger as time goes on; same is true for the oscilation of pollutant concentration. What is worse is that the service rate suffers frequent reduction due to the unstable traffic, which results in traffic jam near the entrance of the links and exacerbates quickly over time (see Figure 2.18). Compared to the lower two plots, the upper two plots in Figure 2.17 show that under the provision of $PM_{2.5}$ concentration information both queues converge to equilibrium smoothly without congestion near the link entrance. The benefit from the provision of air pollution information can therefore be tremendous, as we aim to illustrate and argue in this study.

Finally, we also study scenarios by doubling and halving the WTP parameters with respect to reduction of $PM_{2.5}$ concentration. That is, we set $\beta_c = 0.28$ \$/(μg/m^3) or $\beta_c = 0.07$ \$/(μg/m^3), and keep Δ_d at the base level of 5 min. Figure 2.19 shows the results of total number of cars and $PM_{2.5}$ concentrations on two links. We observe from the upper two plots that in the case with larger β_c, the traffic densities and pollutant concentrations on two links converge to the equilibrium faster than that when $\beta_c = 0.14$ \$/(μg/m^3) (the base case). When β_c is much smaller $\beta_c = 0.07$ \$/(μg/m^3), we can see from the lower

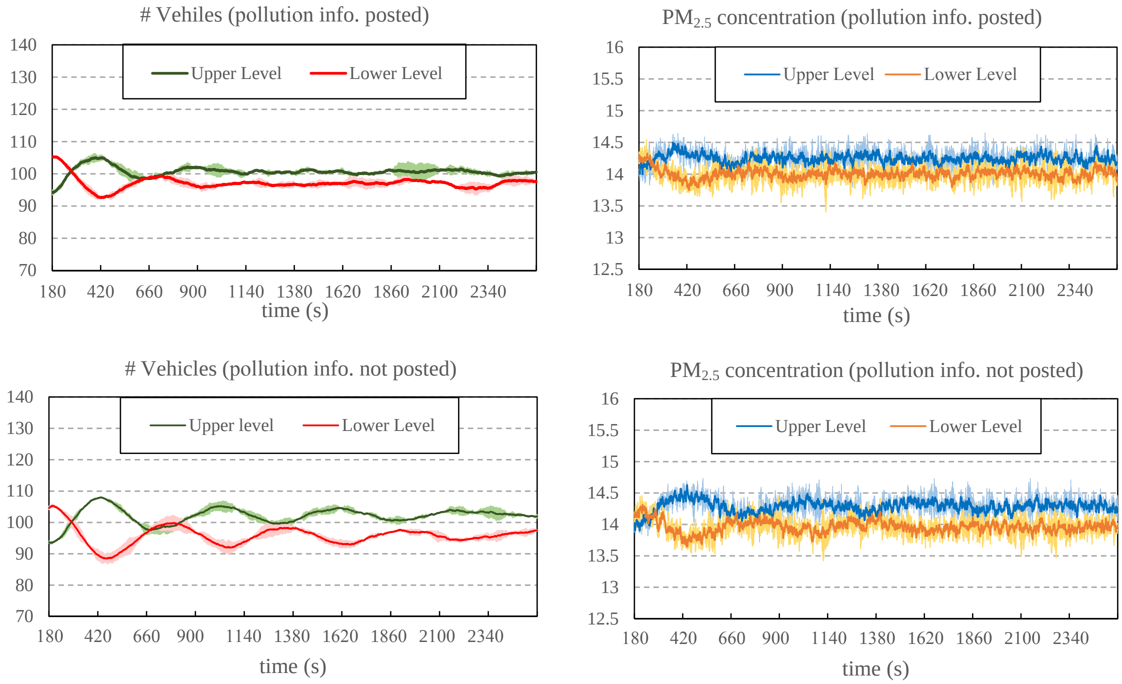


Figure 2.16: Total number of vehicles and pollutant concentration on two links under delay $\Delta = 3$ (min). The dark center line on each curve represents the average of 10^3 independent simulation runs; the lighter-colored area surrounding the dark line on each curve represents the 95% confidence interval for the mean estimate.

two plots that the two queues also converge to stability, although at a slower pace compared to that when β_c is larger. If $\beta_c = 0$ (i.e., there is no provision of air pollution information), the system keeps oscillating with formation of serious congestion and traffic jams at link entrances. Therefore, there exists great potential for traffic improvement and congestion mitigation by provision of air pollution information to drivers as long as travelers attach value, even if it is low, to such information. This supports the discussion we had in the main paper about the effect of β_c on system stability, and it is consistent with results from our analysis of the analytical queueing model (see Theorem 2.2).

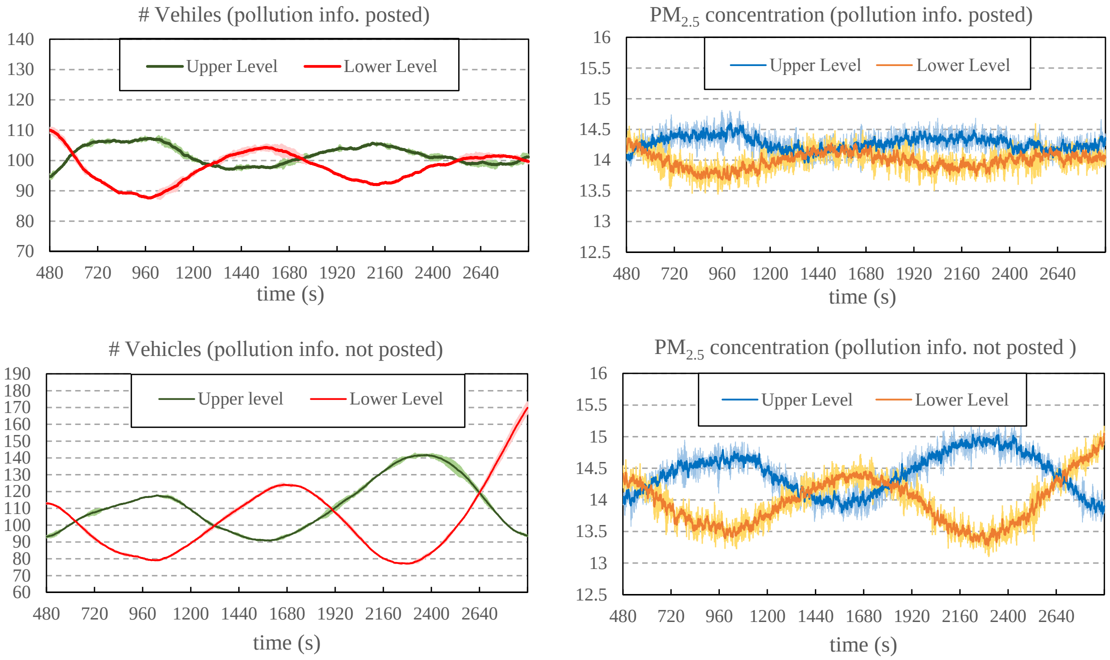


Figure 2.17: Total number of vehicles and pollutant concentration on two links under delay $\Delta = 8$ (min). The dark center line on each curve represents the average of 10^3 independent simulation runs; the lighter-colored area surrounding the dark line on each curve represents the 95% confidence interval for the mean estimate.

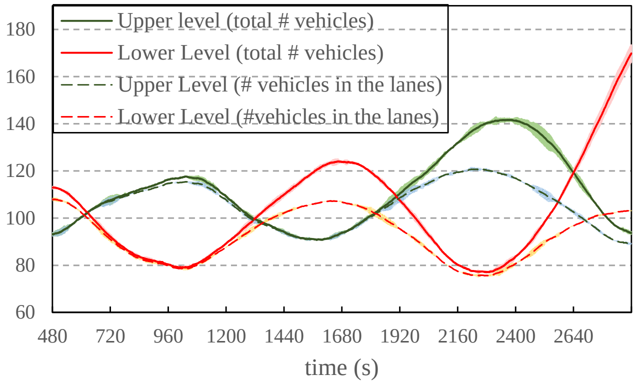


Figure 2.18: Total number of vehicles (including those waiting) and number of vehicles within the lanes when PM$_{2.5}$ concentration information is not posted and $\Delta = 8$ (min). The dark center line on each curve represents the average of 10^3 independent simulation runs; the lighter-colored area surrounding the dark line on each curve represents the 95% confidence interval for the mean estimate.

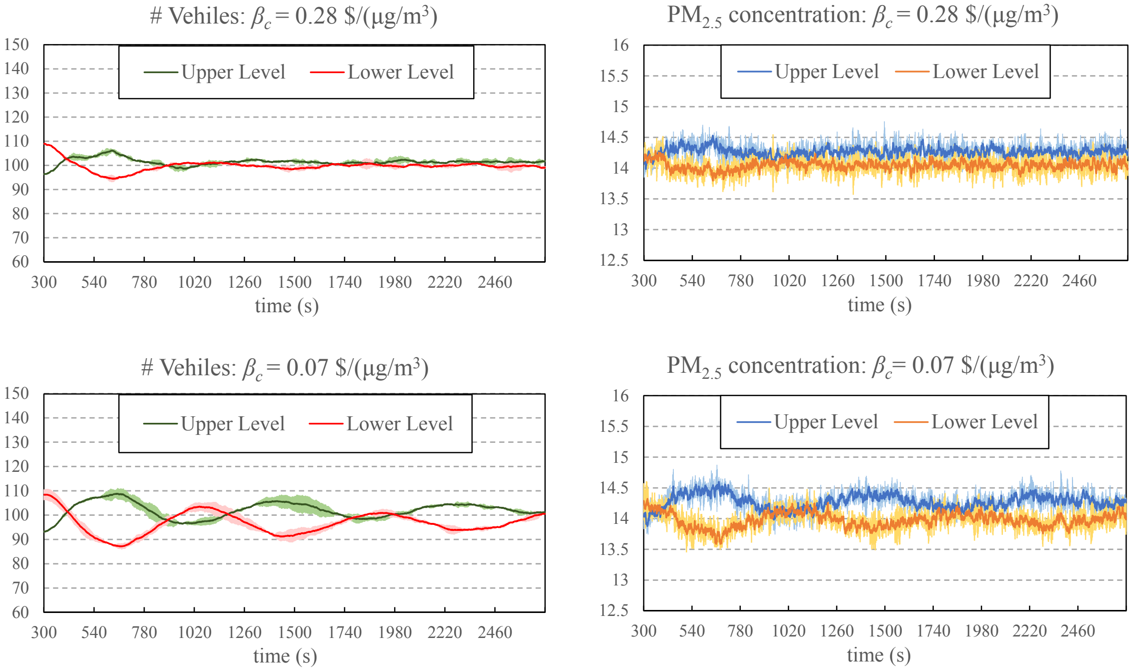


Figure 2.19: Total number of vehicles and pollutant concentration on two links: larger $\beta_c = 0.28$ \$/(μg/m^3) (upper two plots) and $\beta_c = 0.07$ \$/(μg/m^3) (lower two plots). The dark center line on each curve represents the average of 10^3 independent simulation runs; the lighter-colored area surrounding the dark line on each curve represents the 95% confidence interval for the mean estimate.

CHAPTER 3
HYBRID PREDICTIVE CONTROL BASED DYNAMIC PRICING OF MANAGED LANES WITH MULTIPLE ACCESSES

In this chapter, we focus on how new system data and demand information can be fully used and properly incorporated into the model-based control framework for efficient real-time solutions to optimal dynamic traffic management.

For illustration, we propose a hybrid model predictive control (MPC) based dynamic pricing strategy for high-occupancy toll (HOT) lanes with multiple accesses. This approach pre-plans and coordinates the prices for different OD pairs and enables adaptive utilization of HOT lanes by considering available demand information and boundary conditions. It also addresses such practical issues as prevention of recurrent congestion in HOT lanes, ensuring no higher toll for a closer toll exit and fairness among different OD groups at each toll entry, as well as the fact that high occupancy vehicles (HOVs) have free access to the HOT lanes. Taking the inflows at each toll entry as the control, traffic densities and vehicle queue length as observed system states, and boundary traffic as predicted exogenous input, we formulate a discrete-time piecewise affine traffic model. Optimal tolls are then derived from a one- to-one mapping based on the optimal toll entry flows. By properly formulating the constraints, we show that the MPC problem at each stage is a mixed-integer linear program and admits an explicit control law derived by multi-parametric programing techniques. A numerical experiment is presented for a representative freeway segment to validate the effectiveness of the proposed approach. The results show that our control model can react to demand and boundary condition changes by adjusting and coordinating tolls smoothly at adjacent toll entries and drive the system to a new equilibrium that minimizes the total person delay. Under the optimal prediction horizon, the on-line computational cost of the proposed control model is only about 4% and 8% of the modeling cycle of 30 s, respectively, for two typical traffic scenarios, which implies a potential of real-time implementation.

3.1 Introduction

3.1.1 Literature review and motivation

High-occupancy toll (HOT) lanes have been recognized as one of the most applicable and cost-effective measures for reducing freeway congestion [81, 124]. By allowing single-occupancy vehicles (SOVs) to use high-occupancy vehicle (HOV)/carpool lanes by paying a toll, excess capacity of the HOV lanes can be utilized [124]. HOT policy can also help alleviate traffic-related environmental problems and support sustainable urban development [15]. There are three major tolling schemes used in HOT management: static pricing, time-of-day pricing, and dynamic pricing [28, 132]. Static and time-of-day tolls do not reflect or respond to real-time traffic conditions, while dynamic pricing is designed to adjust toll rates according to traffic conditions [28]. However, the performance of dynamic tolling relies on the toll-control algorithm used, which should properly account for traffic information and be implemented efficiently in an on-line fashion. There are various types of toll mechanisms: pass based (i.e., vehicles with a prepaid toll pass can enter the toll lane at any time) , per-use based, per-mile (distance) based, zone (section) based, origin-destination (OD) based, and toll-entry based, and the combination of any of these mechanisms. The first two mechanisms are for single entry/exit managed lanes, while the rest are for systems with multiple toll en- tries/exits (details can be found in [87, 132]). The algorithm we propose in this study is for OD based tolling, which is an ideal toll mechanism for full utilization of the HOT lanes without creating excessive inequality across different OD pairs [87].

A number of HOT lane facilities have been implemented in the U.S., such as on routes I-5, I-10 W, I-15, I-95, I-394, and SR-167 [28, 124, 132], and others are in progress. According to reviews [28, 132], most of these projects use dynamic tolling, and several use distance-based rates [28]. For example, tolls for the I-394 HOT lanes are adjusted every 3 minutes according to the detected traffic density; the toll rates, which are given in a "delta densitytoll increment lookup table" [28, 60, 134] vary from $0.25 to $8.00 across different toll sections and are prescribed to maintain a free-flow traffic speed in the HOT lanes. Similarly, tolls for I-15 change every 6 minutes and vary from $0.5 to $8.00, as given in a lookup table that specifies tolls for different traffic volumes and levels of service (LOS) [28, 69, 132]. Further details of tolling methods can be found in studies such as [28, 132]. These predefined toll adjustment rules often have many parameters to be decided [28]. The introductory period for settling on the rules and determining the parameters can extend to years [42], and the performance of the resulting schemes is still uncertain in future scenarios. Most studies on dynamic tolling have proposed similar elementary rule based models [47]. For example, study [124] proposed a feedback control dynamic tolling algorithm that uses traffic speed as the feedback variable to maintain a high LOS in the HOT lanes and maximize the total throughput; study [134] also proposed a feedback approach that adjusts the toll based on the measured traffic density to maintain the desired traffic density via a linear regulator. Study [49] designed a more complicated approach that optimizes the current toll for full utilization of the HOT lanes based on current demand and the value of time (VOT) distribution.

In reality, speed fluctuations and recurrent congestion are common in managed-lanes, even when there is decent capacity in the system [47, 81]. One important reason for this is a lack of good use of available upstream traffic demand information and measured boundary traffic conditions; toll adjustments are based only on current traffic conditions [47], as is the case in the aforementioned studies. An approach that uses "self-learned" willingness to pay (WTP) parameters was proposed in [134] to derive the toll rate for the next step by solving a one-stage nonlinear optimization problem. Their model uses upstream demand information to predict the traffic in the next step. Study [47] proposed a feedback control approach that also incorporates upstream traffic information and was shown to have a faster response to real-time traffic changes than the simple feedback approach. In both studies, traffic evolution beyond the next step was not taken into account in deriving the tolls; as a result, the performance over a longer period could be unsatisfactory. The "self-learning" approach [134] was extended by [82] to a rolling-horizon setting that considers a longer period of future traffic evolution. However, it optimizes only one future toll input at each stage, which limits its pre-planning capability. In addition, the optimization model is hard to solve, although it considers only a toll range constraint. Study [120] proposed a rolling horizon approach that optimizes a sequence of future tolls with predicted demand from traffic simulation. It solves the nonconvex control problem by exhaustive search. Since the number of feasible tolls is exponential in the horizon length, the real-time applicability of this approach is limited.

All of the aforementioned studies focus on a single toll section. However, real-world HOT projects often have multiple toll entries and exits [43, 132], and new projects tend to have multiple access points with more complicated pricing schemes [43]. For example, there are six main toll entries and exits on the northbound I-15 in San Diego, CA [3], and five main toll entries and exits on the Northeast Loop 820 in Farmers Branch, TX [4]. Research on control in the case of multi-access managed lanes is relatively limited. The authors of [47] extended their single-toll-entry control method to a multi-access system but proposed only a very simple heuristic with no practical constraints. A general simulation model for HOT lanes with multiple toll sections was developed in [87] that can be tailored to various toll mechanisms. Study [36] proposed a dynamic tolling model that first designs the flow split ratios to the two types of lanes and then sets the tolls based on VOT distributions or auctions. However, it uses no information about incoming flows. A distance-based dynamic tolling model was developed in [132] for managed lanes with multiple entries and exits. It uses a quite realistic nonlinear continuous-time traffic-flow model but makes the control problem nonconvex and highly intractable. The work [138] proposed an interesting reinforcement-learning (RL) based approach for dynamic per-mile tolling of managed lanes with flexible accesses. Real-time computation seems to be a limiting factor and the model considers only a few traffic densities and toll rates due to "curse of dimensionality.

The aforementioned issues need to be considered and addressed in the design of practical and more effective HOT-lane management strategies. In this study we formulate the managed lane system as a tractable hybrid system together with a model predictive control (MPC) based tolling strategy. Our model has two key features: 1) it can handle systems with multiple HOT-lane accesses by optimizing tolls for each toll entry-exit pair utilizing the available upstream demand information; 2) it has the flexibility to accommodate various constraints (such as free-flow on HOT lanes) while admits convenient real- time implementation. The objective of the control model is to minimize the total person delay in the system, thus improving the social welfare. We propose a proper "fairness" condition among OD pairs for each toll entry without affecting capacity utilization of HOT lanes, and the constraint is compatible with the restriction of no higher tolls for closer toll exits. MPC framework has been proposed for other traffic control measures, such as variable speed limits, ramp metering, or a combination of the two, and was shown to be an effective tool [12, 55, 62, 83]. However, if the tolls are optimized directly (as in [49, 82, 120, 132]), the control model can be very complicated due to the nonconvex lane-choice probability functions. We instead take the toll entry flows as the control input and show that the optimal tolls can be derived from those flows. By properly formulating the constraints, we prove that the control problem is a mixed-integer linear program (MILP) that admits explicit control law based on multi-parametric programing techniques. Note that our hybrid system model formulation and the two-step toll design approach overcome the computational intractability of the models used in previous studies (e.g., [82, 132]). Moreover, compared to the RL approach [138], our method solves for the optimal tolls and controls the traffic densities both in continuous space without limiting the optimality of the solution.

3.1.2 Model assumptions and notations

Assumption 3.1 *Automatic electronic toll facilities are used such as the "transponders" [3]. The SOVs traveling to the same exit pay the same toll rate they see while entering the toll entry, this is imposed by sensors that record where each vehicle enters and exits the HOT lane and toll the trip. Thus the proposed tolling approach is OD-based, which is an ideal toll scheme [87]. Vehicle occupancy, traffic density and volume are measured. Upstream demand forecast has relatively high accuracy within the prediction horizon.*

Assumption 3.2 *The OD ratios (i.e., the flow proportions for individual toll entryexit pairs) of boundary inflows are the same for the SOVs as for the HOVs, and are unchanged within the problem horizon, so do the HOV proportions at each boundary inflow. Traffic demand of the OD pairs is such that the recurrent-congestion-free constraint for HOT lanes can be satisfied and the toll entry/exit flows can be accepted under proper tolls. However GP lanes can be congested.*

Assumption 3.3 *A HOV prefers HOT lanes to GP lanes. The lane-choice probabilities of an SOV can be described by a Logit model [124, 134, 138] that has linear utility function with constant parameters in the problem horizon. The utility comes from travel time, toll and other factors. The other factors (e.g., travel time variation) give no higher utility traveling in GP lanes than in HOT lanes [24] and this utility difference is increasing in travel distance.*

Assumption 3.4 *The SOVs only consider to enter the first toll entry they encounter where their destination is at least as far as the next toll exit, and leave the HOT lanes (if entered) when approach their target exit of the freeway segment. This behavioral* *assumption is mainly for practical toll control algorithm design considerations and is also implicitly used in relevant studies (e.g., [36]). It is also consistent with Assumption 3.2 since otherwise the potential demand for each toll entry can be toll-dependent. This assumption also automatically holds if the toll entries (from the second one) are all direct HOT accesses.*

Main notations used in our model are listed in Table 3.1.

The rest of this chapter is organized as follows. The next section presents models describing traffic flows and lane choices for a multi-access managed-lane system. Section 3.3 discusses the toll-manipulation model with practical constraints. In Section 3.4, the control model is formulated and analyzed. A numerical experiment is presented in Section 3.5. Finally, in Section 3.6 we draw conclusions and outline future research.

3.2 Modeling of Multi-Access Managed-Lane System

3.2.1 Traffic-flow model

Traffic dynamics on a freeway segment with managed lanes can be modeled using the popular macroscopic traffic dynamic model: the cell transmission model (CTM) [32]. We extend the idea of cell partition proposed in the original CTM to managed lane systems. Specifically, we consider major bypass inflows and divide the mainline freeway segment into N cell pairs governed by the following principles (similar to [36, 132]), the length can vary across different cell pairs; 2) each cell pair contains at most one toll entry (near its start), either at a bypass

Table 3.1: Notation

n	Number of HOT/GP cell pairs on the freeway segment
L_i	Length of cell pair i (mile)
$v_{f,i}^H/v_{f,i}^G$	Normalized free-flow speed in HOT lanes / GP lanes in cell pair i
n_i^H/n_i^G	Number of vehicles in HOT lanes / GP lanes in cell pair i
q_i^H/q_i^G	Number of vehicles leaving cell i and moving to cell $i+1$ in HOT lanes / GP lanes
$n_{c,i}^H/n_{c,i}^G$	Critical density of HOT lanes / GP lanes in cell pair i
$n_{J,i}^H/n_{J,i}^G$	Jam density of HOT lanes / GP lanes in cell pair i
w_i^H/w_i^G	Normalized congestion wave speed in HOT lanes / GP lanes in cell pair i
$q_{M,i}^H/q_{M,i}^G$	Flow capacity in HOT lanes / GP lanes from cell i to $i+1$
v_i^H/v_i^G	Travel speed in HOT lanes / GP lanes in cell pair i (mile/h)
l	Number of vehicles waiting to enter the first GP lane cell
d_{ij}	Travel distance from (the start of) cell pair i to (the end of) cell pair j (mile)
$\tau_{ij}/\tau_{ij}^{\max}$	Toll / maximum toll for travel in HOT lanes from cell pair i to cell pair j ($)
t_{ij}^H/t_{ij}^G	Perceived travel time from cell pair i to cell pair j in HOT lanes / GP lanes (h)
f_i^{in}/f_i^{out}	Traffic flow that enters/ exits HOT lanes at cell pair i (veh/h)
r_i/η_i	Total inflow upstream the toll entry at cell pair i (veh/h)/proportion of HOVs in r_i
λ_i	proportion of vehicles travelled through GP lane cell i and keep moving to cell $i+1$
α_{1i}/α_{2i}	Marginal utility of travel time / toll for SOVs at the toll entry in cell pair i
γ_{ij}	dis-utility of traveling via d_{ij} via GP lanes compared to HOT lanes due to factors other than time and toll
Δt	Length of the modeling time step (min)
β_{ij}	Proportion of flow among r i that can travel in HOT lanes and exit at cell pair j
p_{ij}^L	Probability that an SOV will choose the HOT lanes and travel from cell pair i to j
n_i^{HOV}/o_i	Number of HOVs / average occupancy of HOVs in cell pair i

line that has on-ramps to both HOT lanes (direct access) and GP lanes or an entry from GP lanes to HOT lanes near an upstream GP lane on-ramp; 3) each cell pair has at most one toll exit (near its end), either a HOT lane off-ramp (in which case this cell pair also has a GP lane off-ramp) or an exit from HOT lanes to GP lanes near a downstream GP lane off-ramp; 4) as required for the convergence of CTM [32], vehicles can travel through no more than one cell in any time step. Figure 3.1 shows a freeway segment with $N = 4$ cell pairs.

In the sequel, we use superscripts "H" and "G" to denote HOT lanes and GP lanes, respectively.

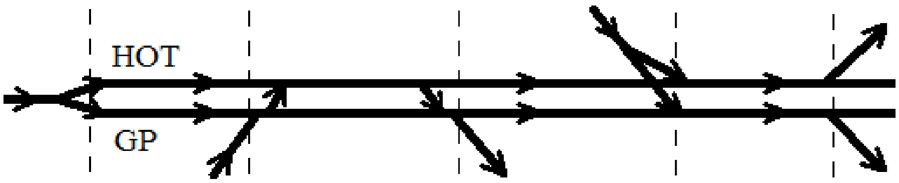


Figure 3.1: Freeway segment partitioned into cells.

Traffic flow dynamics at each freeway cell

In our discrete time model, the time interval is indexed by t and has length Δt. For $i = 1, ..., N$ and each lane type, $q_i(t)$ is the number of vehicles leave cell i and move to cell $i + 1$ during time step t. By the CTM [32, 104], $q_i(t)$ equals to the minimum over three quantities from left to right in (3.1): the number of vehicles that can be sent by cell i to $i + 1$ during time step t, the number of vehicles that can be received by cell $i + 1$ from i , and the maximum possible flow from cell i to $i + 1$:

$$\begin{aligned} q_i^H(t) &= \min[v_{f,i}^H n_i^H(t) - f_i^{out}(t),\ w_i^H(n_J^H - n_{i+1}^H(t)),\ q_{M,i}^H - f_i^{out}(t)], \\ q_i^G(t) &= \min[\lambda_i v_{f,i}^G n_i^G(t),\ w_i^G(n_J^G - n_{i+1}^G(t)),\ \lambda_i q_{M,i}^G], \end{aligned} \tag{3.1}$$

where $n_i(t)$ is the number of vehicles (traffic density) in cell i at time step t; $\lambda_i \in [0, 1]$ is the proportion of vehicles travelled through GP lane cell i and keep moving to GP lane cell $i + 1$, which is assumed constant during the problem horizon; $q_{M,i}$ is the flow capacity, i.e., the maximum number of vehicles can travel from cell i to $i + 1$ in each time step if $\lambda_i = 1$. $f_i^{out}(t)$ is the number of vehicles leave HOT lane cell i in time step t; $v_{f,i}$, $w_i \in (0, 1]$ are the normalized free-flow speed and congestion wave speed of cell i, respectively; $n_{c,i}$, $n_{J,i}$ are the critical density and jam density of cell i , respectively.

The goal for the managed-lane system is to keep free-flow traffic in HOT lanes [28], so we restrict $0 \leq n_i^H \leq n_{c,i}^H$ for all $i = 1, ..., n$. Then (3.1) for HOT lane cells simplifies to

$$q_i^H(t) = v_{f,i}^H n_i^H(t) - f_i^{out}(t),\ i = 1, ..., N. \tag{3.2}$$

We assume that the bypass inflow r i at downstream GP lane cell i ($i > 1$) is notably smaller than the demand for the first toll entry of the modeling segment, so we do not consider ramp metering at cell i ($i > 1$). However, the accumulating bypass inflows (due to the absence of ramp metering) can result in congestion in GP lanes, a natural consequence is the formation of a vehicle queue in the first cell of the modeling segment [53]. We model this effect by maintain a vehicle queue in front of the first GP lane cell. Hence for the first cell of each lane type, we have

$$q_0^H(t) = f_1^{in}(t),\ q_0^G(t) = \min[l(t) + r_1(t) - f_1^{in}(t),\ w_1^G(n_{J,1}^G - n_1^G(t)),\ q_{M,0}^G] \tag{3.3}$$

where r_1 is the demand upstream mainline entry; l is the length of the vehicle queue in front of the first GP lane cell; f_i^{in} is the number of vehicles enter HOT lane cell i.

Since GP lanes can be congested, i.e., only $0 \leq n_i^G \leq n_{J,i}^G$ needs to hold for all $i = 1, ..., N$, then (3.1) and (3.3) for each GP lane cell can be rewritten as (time index omitted for clarity)

$$\begin{aligned}
q_0^G &= \begin{cases} l + r_1 - f_1^{in} & \text{if } l + r_1 - f_1^{in} \leq \min[w_1^G(n_{J,1}^G - n_1^G),\ q_{M,0}^G] \\ w_1^G(n_{J,1}^G - n_1^G) & \text{if } w_1^G(n_{J,1}^G - n_1^G) \leq \min[l + r_1 - f_1^{in},\ q_{M,0}^G] \\ q_{M,0}^G & \text{if } q_{M,0}^G \leq \min[l + r_1 - f_1^{in},\ w_1^G(n_{J,1}^G - n_1^G)] \end{cases}, \\
q_i^G &= \begin{cases} \lambda_i v_{f,i}^G n_i^G & \text{if } \lambda_i v_{f,i}^G n_i^G \leq \min[w_{i+1}^G(n_{J,i+1}^G - n_{i+1}^G),\ \lambda_i q_{M,i}^G] \\ w_{i+1}^G(n_{J,i+1}^G - n_{i+1}^G) & \text{if } w_{i+1}^G(n_{J,i+1}^G - n_{i+1}^G) \leq \min[\lambda_i v_{f,i}^G n_i^G,\ \lambda_i q_{M,i}^G] \\ \lambda_i q_{M,i}^G & \text{if } \lambda_i q_{M,i}^G \leq \min[\lambda_i v_{f,i}^G n_i^G,\ w_{i+1}^G(n_{J,i+1}^G - n_{i+1}^G)] \end{cases}, \quad i = 1, ..., N-1, \\
q_N^G &= \begin{cases} \lambda_N v_{f,N}^G n_N^G & \text{if } \bar{q}_N^G \geq v_{f,N}^G n_N^G \\ \lambda_i \bar{q}_N^G & \text{if } \bar{q}_N^G \leq v_{f,N}^G n_N^G \end{cases},
\end{aligned} \tag{3.4}$$

where we define $\bar{q}_N^G = \min[\hat{w}_{N+1}^G(n_{J,N+1}^G - \hat{n}_{N+1}^G),\ \lambda_N q_{M,N}^G]$, which can be determined by measured density $\hat{n}_{N+1}^G$ and congestion wave speed $\hat{w}_{N+1}^G$ downstream the modeling segment. Note that model (3.1)-(3.4) involves toll entry and exit flows, f_i^{in} and f_i^{out}, queue length l, and proportions of retaining flows, λ_i. These quantities are not present in the original CTM [32], which was for a simple freeway segment (one lane type, no bypass inflows/outflows). However, derivation of our model (3.1)-(3.4) inherits the basic idea of the original CTM and extends it to the managed lane system. We also have the following two remarks on the model above.

Remark 3.1 *We assumed that the demand upstream the modeling segment is notably larger than the bypass inflows and thus we only maintain a vehicle queue upstream the first GP lane cell of the modeling segment. In practice if the system is very long with several relatively large bypass inflows (e.g., from a connector of another freeway), we can first divide the entire system into several modeling segments each of which starts* *at the location of a dominant inflows. Then using proper priority rules at the freeway junctions involving such dominant inflows [36], the hybrid model we proposed later can be extended to multi-segment systems and a distributed control framework can be adopted. As we focus on the tolling problem in a representative freeway segment, we leave this extension as a future study.*

Remark 3.2 *The use of constant proportion λ_i (defined in (3.1)) is similar to the use of off-ramp splitting ratio in the freeway control literature (e.g., [53, 91, 104]). Strictly speaking, λ_i depends not only on the OD ratios but also on the relative relationship among traffic demand approaching different toll entries upstream of cell pair i and the flows entering these toll entries (and thus λ_i also depends on the tolls at these upstream toll entries). As shown in [41], modeling the off-ramp flow as a function of inflows from upstream on-ramps makes dynamic ramp metering problem intractable. The tolling problem is more complex than ramp metering, so we use constant λ_i's for model tractability and practicability. However, we can actually impose proper constraints on the flow proportions entering each toll entry to improve the accuracy of the approximation by constant λ_i (see Section 3.3.2 for details).*

By the conservation of vehicles, we have the dynamics of vehicle densities and queue length as

$$n_i^H(t+1) = n_i^H(t) + q_{i-1}^H(t) - q_i^H(t) + f_i^{in}(t) - f_i^{out}(t),\ i = 1, ..., N, \tag{3.5}$$

$$n_i^G(t+1) = \begin{cases} n_i^G(t) + q_{i-1}^G(t) - q_i^G(t)/\lambda_i & \text{if } i = 1 \\ n_i^G(t) + q_{i-1}^G(t) - q_i^G(t)/\lambda_i + r_i(t) - f_i^{in}(t) & \text{if } i = 2, ..., N \end{cases}, \tag{3.6}$$

$$l(t+1) = l(t) + r_1(t) - f_1^{in}(t) - q_0^G(t). \tag{3.7}$$

Traffic dynamics of the entire system

The complexity of the traffic dynamics mainly comes from undetermined congestion status on GP lane cells. Thus in this subsection we focus on explicitly modeling of the dependency between density evolution in the system to the congestion status on GP lanes.

For easy of notation, we define $\lambda_0 = v^G_{f,0} = w^G_{N+1} = 1$, $n^G_0 = l + r_1 - f^{in}_1$, $n^G_{N+1} = n^G_{J,N+1}\bar{q}^G_N$. Let vectors $n^H = [n^H_1, ..., n^H_N]^{\mathrm{T}}$, $n^G = [n^G_0, ..., n^G_{N+1}]^{\mathrm{T}}$, $f^{in} = [f^{in}_1., ..., f^{in}_N]^{\mathrm{T}}$, $r = [r_1, ..., r_N]^{\mathrm{T}}$. By the bounds on each entry of n_G, we know that $n_G \in \Omega = [0, q^G_{M,0}]\times[0, n^G_{J,1}]\times...\times[0, n^G_{J,N}]\times[0, n^G_{J,N+1}] \subset \mathbb{R}^{N+2}$. We then define a set of polyhedra

$$\begin{aligned} D_i &= \left\{n^G \in \Omega | \mu^{(1)}_i n^+_i \le \zeta^{(1)}_i,\ \mu^{(2)}_i n^+_i \le \zeta^{(2)}_i\right\},\ i = 0, ..., N, \\ W_i &= \left\{n^G \in \Omega | \mu^{(1)}_i n^+_i \ge \zeta^{(1)}_i,\ \mu^{(3)}_i n^+_i \le \zeta^{(3)}_i\right\},\ i = 0, ..., N, \\ L_i &= \left\{n^G \in \Omega | \mu^{(2)}_i n^+_i \ge \zeta^{(2)}_i,\ \mu^{(3)}_i n^+_i \ge \zeta^{(3)}_i\right\},\ i = 0, ..., N, \end{aligned} \tag{3.8}$$

where $n^+_i = [n^G_i, n^G_{i+1}]^{\mathrm{T}}$, $\mu^{(j)}_i \in \mathbb{R}^3$, $\zeta^{(j)}_i \in \mathbb{R}$, $i = 0, ..., N$, $j = 1, 2, 3$ are the constant coefficients

$$\begin{aligned} &\mu^{(1)}_i = [\lambda_i v^G_{f,i}\ w^G_{i+1}],\ \mu^{(2)}_i = [\lambda_i v^G_{f,i}\ 0],\ \mu^{(3)}_i = [0\ \ -w^G_{i+1}], \\ &\zeta^{(1)}_i = w^G_{i+1} n^G_{J,i+1},\ \zeta^{(2)}_i = \lambda_i q^G_{M,i},\ \zeta^{(3)}_i = \lambda_i q^G_{M,i} - w^G_{i+1} n^G_{J,i+1}. \end{aligned} \tag{3.9}$$

Then we have a basic result regarding the flow dynamics in the GP lanes.

Lemma 3.1 *The traffic density at GP lane cells can be described as*

$$n^G_1(t+1) = F_1(m_1)[n^G_0(t),\ n^G_1(t),\ n^G_2(t)]^{\mathrm{T}} + a_1(m_1), \tag{3.10}$$

$$n^G_i(t+1) = F_i(m_i)[n^G_{i-1}(t),\ n^G_i(t),\ n^G_{i+1}(t)]^{\mathrm{T}} + a_i(m_i) + r_i(t) - f^{in}_i(t),\ i = 2, ..., N,$$

where $m_i \in \Pi = \{1, ..., 9\}$ *represents the possible modes of each GP lane cell and* $F_i(\cdot) :\to \mathbb{R}^3$, $a_i(\cdot) :\to \mathbb{R}$ *are cell-mode-dependent coefficients (given in Table 3.2),* $i = 1, ..., N$.

Proof: Substituting (3.4) into (3.6) and by the polyhedra defined in (3.8)-(3.9), we obtain six different linear dynamics for n_1 and n_N, and nine different linear dynamics for $n_i,\ i = 2, ..., N-1$, at different polyhedra. The coefficients of these linear dynamics are shown in Table 3.2. ■

Table 3.2: Coefficients in (3.10) under different modes (superscript "G" omitted)

m_i	$n^+_{i-1} \in$	$n^+_i \in$	$F_i(m_i)$	$a_i(m_i)$
1^a	L_{i-1}	L_i	$[0,\ 1,\ 0]$	$\lambda_{i-1}q_{M,i} - q_{M,i}$
2	L_{i-1}	W_i	$[0,\ 1,\ w_{i+1}/\lambda_i]$	$\lambda_{i-1}q_{M,i} - w_{i+1}n_{J,i+1}/\lambda_i$
3	L_{i-1}	D_i	$[0,\ 1 - v_{f,i},\ 0]$	$\lambda_{i-1}q_{M,i}$
4^a	W_{i-1}	L_i	$[0,\ 1 - w_i,\ 0]$	$w_i n_{J,i} - q_{M,i}$
5	W_{i-1}	W_i	$[0,\ 1 - w_i,\ w_{i+1}/\lambda_i]$	$w_i n_{J,i} - w_{i+1}n_{J,i+1}/\lambda_i$
6	W_{i-1}	D_i	$[0, 1 - v_{f,i} - w_i,\ 0]$	$w_i n_{J,i}$
7^a	D_{i-1}	L_i	$[\lambda_{i-1}v_{f,i-1},\ 1,\ 0]$	$-q_{M,i}$
8	D_{i-1}	W_i	$[\lambda_{i-1}v_{f,i-1},\ 1,\ w_{i+1}/\lambda_i]$	$-w_{i+1}n_{J,i+1}/\lambda_i$
9	D_{i-1}	D_i	$[\lambda_{i-1}v_{f,i-1},\ 1 - v_i,\ 0]$	0

[a] These m_i's are impossible for $i = N$ since $L_N = \emptyset$ by construction.

Define $m = [m_1, ..., m_N]^{\mathrm{T}}$ be the mode vector of the entire GP lane segment. It seems that m can take possibly $6 \times 9^{N-1}$ many values. However, this number can be significantly reduced as the modes of two adjacent cell pairs are correlated, e.g. if $m_{i-1} = 1$ then m_i can only be 1, 2 or 3.

Lemma 3.2 *Let M be the total number of possible mode vectors m, then $M = 2^N$.*

Proof: From Table 3.2 we see that the relevant components of n^G can be in any of the sets W_i, D_i or L_i for $0 \leq i < N$ and can be in any of sets W_N or D_N, so by definition, the total number of different mode m is 2×3^N. ■

Remark 3.3 *If a triangular fundamental diagram (FD) is used (i.e., $v_i n_{c,i} = w_i(n_{J,i} - n_{c,i}) = q_{M,i}$) and $q_{M,i} = q_M$, we can show that mode $m_i = 6$ can be removed. If in addition $\lambda_{i-1} = 1$, then $m_i = 1$ can also be removed, in which case it is proved in [118] that M grows in the order of about 2.25^N asymptotically. Notice that our result (Lemma 3.2) is more general, since it holds for freeway cells with heterogeneous parameters and any plausible shape of FDs.*

Let $w = [r^{\mathrm{T}}, n^G_{N+1}]^{\mathrm{T}}$ be the vector of exogenous input. By (3.8)-(3.9) and Table 3.2, we can define the following set of polytopes that form a partition of Ω

$$\Omega_k = \left\{n^G \in \Omega : \text{mode vector is } m_k\right\} = \left\{n^G \in \Omega : D_k n^G \leq d_k\right\}, \ k = 1, ..., M. \qquad (3.11)$$

Then we can write (3.10) as a piecewise affine (PWA) system:

$$n^G(t+1) = A^G_k n^G(t) + B^G w(t) + C^G f^{in}(t) + a^G_k, \text{ if } \delta_k = 1, \ k = 1, ..., M.$$

$$A^G_k = \begin{bmatrix} \mathbf{0}^{\mathrm{T}} & \dots & \mathbf{0}^{\mathrm{T}} \\ F_1(m^k_1) & & \\ & \ddots & \\ & & F_N(m^k_N) \\ \mathbf{0}^{\mathrm{T}} & \dots & \mathbf{0}^{\mathrm{T}} \end{bmatrix}, \ B^G = \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ 0 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & I \end{bmatrix}, \ C^G = \begin{bmatrix} -e^{\mathrm{T}}_1 \\ \mathbf{0}^{\mathrm{T}} \\ -I \\ \mathbf{0}^{\mathrm{T}} \end{bmatrix}, \qquad (3.12)$$

where $m_k = [m^k_1, ..., m^k_N]^{\mathrm{T}}$ is the k^{th} mode vector and we define $\delta_k(t) \in \{0, 1\}$, $k = 1, ..., M$, satisfying $\delta_k(t) = 1 \Leftrightarrow n^G(t) \in \Omega_k$ and $\sum_k \delta_k(t) = 1$. Recall that the first and last entries in $n^G \in \mathbb{R}^{N+2}$ represent the auxiliary densities upstream and downstream the modeling segment: $n^G_0 = l + r_1 - f^{in}_1$, $n^G_{N+1} = n^G_{J,N+1} - \bar{q}^G_N$, thus the first and last rows in A^G_k, B^G and C^G take the values above (e_1 is a N-dim vector with first entry one and the others zero, and I is a N-by-N identity matrix).

Lemma 3.3 *System (3.12) is well-posed, i.e., for any $n^G(t) \in \Omega$, the mapping $(n^G(t),\ w(t),\ f^{in}(t)) \to n^G(t+1)$ defined by (3.12) is single valued.*

Proof: By definition, n_i^+ $(i = 0, ..., N)$ are subvectors of $n^G(t) \in \Omega$, then considering (3), we know that the polyhedra $1, ..., M$ are all closed and have disjoint interior with their union being the polytope. Thus for any $n^G(t) \in \Omega$, either there exists only one index $k \in \{1, ..., M\}$ satisfying $n^G(t) \in \Omega_k$, or there exits more than one indices $k \in \{1, ..., M\}$ with $n^G(t)$ lies on the common boundary of these k's. Since the mapping $(n^G(t),\ w(t),\ f^{in}(t)) \rightarrow n^G(t+1)$ is continuous on its domain, so it is single valued. ■

Now we have the following crucial reformulation of the system model (3.12) as a set of linear constraints on mixed integer variables.

Theorem 3.1 *Let* $\bar{n} = [q^G_{M,0}, n^G_{J,1}, ..., n^G_{J,N+1}]^{\mathrm{T}}$*, (3.12) is equivalent to the following constraints*

$$\begin{aligned}
&n^G(t+1) = \textstyle\sum_{k=1}^M s_k(t),\ \delta_k(t) \in \{0, 1\},\ \sum_{k=1}^M \delta_k(t) = 1,\\
&g_k[1 - \delta_k(t)] \geq D_k n^G(t) - d_k,\ \bar{n}\delta_k(t) \geq s_k(t) \geq \mathbf{0},\\
&s_k(t) \leq A^G_k n^G(t) + B^G_k w(t) - f^{in}(t),\\
&s_k(t) \geq A^G_k n^G(t) + B^G_k w(t) - f^{in}(t) - \bar{n}[1 - \delta_k(t)],
\end{aligned} \tag{3.13}$$

where $g_k = \max_{n^G \in \Omega} D_k n^G - d_k,\ k = 1, ..., M.$

Proof: Clearly, for any $k = 1, ..., M$, $\mathbf{0} \leq A^G_k n^G + B^G w - f^{in} \leq \bar{n}$ (entry-wise comparison) holds, and for either the upper or lower bound, there must exit one k such that the bound is tight by our partition of into $\{k, k = 1, ..., M\}$. Thus the result follows by Lemma 3 and the discussion in [13]. ■

For the dynamics of HOT lane densities in (3.5), we simply have

$$n^H(t+1) \quad = \quad A^H n^H(t) + f^{in}(t), \tag{3.14}$$

$$A^H = \begin{bmatrix} 1-v_1^H & 0 & \dots & 0 & 0 \\ v_1^H & 1-v_2^H & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1-v_{N-1}^H & 0 \\ 0 & 0 & \dots & v_{N-1}^H & 1-v_N^H \end{bmatrix}$$

Note that the toll exit flow terms $\{f_i^{out}\}$ do not appear in (3.14), this is because they are cancelled by (3.2) and (3.5) as a result of free-flow traffic in HOT lanes.

Finally, for the dynamics of the queue length at segment entry, we can express (3.7) as

$$\begin{aligned} l(t+1) &= (1-\sigma_k^D)l(t) + (1-\sigma_k^D)e_1^{\mathrm{T}}[w(t) - f^{in}(t)] + \sigma_k^W w_1^G[e_2^{\mathrm{T}} n^G(t) - n_{J,1}^G] - \sigma_k^L q_{M,0}^G \\ &= l(t) - \sigma_k^D e_1^T n^G(t) + \sigma_k^W w_1^G[e_2^{\mathrm{T}} n^G(t) - n_{J,1}^G] - \sigma_k^L q_{M,0}^G \qquad (3.15) \\ & \begin{cases} \sigma_k^L = 1 & \text{if } m_1^k \in \{1,2,3\} \text{ and 0 otherwise,} \\ \sigma_k^W = 1 & \text{if } m_1^k \in \{4,5,6\} \text{ and 0 otherwise,} \\ \sigma_k^D = 1 & \text{if } m_1^k \in \{7,8,9\} \text{ and 0 otherwise,} \end{cases} \end{aligned}$$

where the second equality follows by our definition $n_0^G = l + r_1 - f_1^{in}$; and the binary variables σ_k^L, σ_k^W, σ_k^D encode if the term $-q_0^G$ in (3.7) is equal to $q_{M,0}^G$, $w_1^G(n_J^G - n_1^G)$ or $l + r_1 - f_1^{in}$, with $\sigma_k^L + \sigma_k^W + \sigma_k^D = 1$, so (3.15) is also a well-posed PWA system.

In practice the queue lengths must be bounded, thus we can impose $l(t) \leq l_{\max}$. Then by noting the correspondence between σ_k^L, σ_k^W, σ_k^D and the binary variables $\delta_k(t)$, $k = 1, ..., M$, we can express (3.15) also as a set of linear relations in terms of continuous variables $l(t)$, $l(t+1)$ and binary variables $\delta_k(t)$, $k = 1, ..., M$, in the similar way as we translated (3.12) to (3.13).

Combining (3.13), (3.14) and (3.15), we can describe the traffic dynamics in the entire managed lane system by a set of linear mixed-integer relations, which is the basis for the control model design.

3.2.2 Lane-choice model

The main factors that affect lane choice behavior of SOVs are the difference in the travel times for the HOT lanes and the GP lanes and the toll; other minor factors also exist [134]. In this study the commonly used linear utility function is adopted. Let d_{ij} be the travel distance from start of cell pair i to end of cell pair j, t^H_{ij} and t^G_{ij} be the perceived travel times in the HOT lanes and the GP lanes, respectively, on this distance. Since we ensure free-flow in the HOT lanes, $t^H_{i,i+1} = d_{i,i+1}/v^H_{f,i}$. However, $t^G_{ij}(t)$ depends on GP lane congestion status. We assume that$t^G_{ij}(t)$ is equal to d_{ij} divided by the average of the historical mean speed users experienced over distance d_{ij} via GP lanes, $\bar{v}^G_{ij}$, and the real-time local speed, $v^G_i(t)$. Hence the utility of an SOV user traveling from cell pair i to cell pair j via the HOT lanes (H) or GP lanes (G) is

$$\begin{cases} U^H_{ij}(t) = \alpha_{1i} t^H_{ij} + \alpha_{2i}\tau_{ij}(t) = \alpha_{1i}\displaystyle\sum_{j'=i}^{j} d_{j',j+1}/v^H_{f,j'} + \alpha_{2i}\tau_{ij}(t), \\ U^G_{ij}(t) = \alpha_{1i} t^G_{ij} + \gamma_{ij} = 2\alpha_{1i} d_{ij}/\left(v^G_i(t) + \bar{v}^G_{ij}\right) + \gamma_{ij} \end{cases} \tag{3.16}$$

where $\tau_{ij}(t)$ is the toll ($) for travel of an SOV in the HOT lanes from cell pair i to cell pair j; an SOV that enters the HOT lanes in cell pair i during time interval t will be charged $\tau_{ij}(t)$ upon exiting the HOT lanes in cell pair j (see Assumption 3.1). The parameters $\alpha_1 \leq 0$, $\alpha_2 < 0$ represent the marginal effects of the travel time and the toll, respectively, on an SOV's travel utility, which can be location dependent, and α_1/α_2 represents an SOV's VOT. γ_{ij} represents the utility due to other factors such as travel time variation in GP lanes relative to HOT lanes [24].

We have $\gamma_{ij} \leq 0$ by Assumption 3.3.

Thus, for each toll entry-exit pair, the probability that an SOV will choose the HOT lanes is:

$$\begin{aligned} P^L_{ij}(t) &= P^L_{ij}(v^G_i(t),\ \tau_{ij}(t)) = [1 + \exp(U^G_i j(t) - U^H_{ij}(t))]^{-1} \\ &= \left\{1 + \exp[\alpha_{1i}(t^G_{ij}(t) - t^H_{ij}) - \alpha_{2i}\tau_{ij}(t) + \gamma_{ij}]\right\}^{-1} \end{aligned} \tag{3.17}$$

We write $P^L_{ij}(t)$ as a function of $v^G_i(t)$ and $\tau_{ij}(t)$ since it depends on t only through $v^G_i(t)$ and $\tau_{ij}(t)$ by (3.16). The model parameters α_1, α_2 and γ_{ij} can be estimated empirically [24], we will give an example in Section 3.5.

Based on the lane-choice probability P^L_{ij} for SOVs by (3.17), the common OD ratios for HOVs and SOVs from Assumption 3.2, and the preference of HOVs for HOT lanes from Assumption 3.3, we can calculate P_{ij}, the expected proportion of the vehicle flow that will enter the HOT lanes via the toll entry in cell pair i and head for the toll exit in cell pair j (this flow is denoted by r_{ij}), to be

$$P_{ij}(t) = (1 - \eta_i(t))P^L_{ij}(t) + \eta_i(t), \tag{3.18}$$

where η_i is the proportion of HOVs among the flow r_i, it can be time variant.

3.2.3 Toll entry flow model

By Assumption 3.4 we know the possible flow which can potentially enter the HOT lane at cell pair i is r_i. In addition, we note that the underlying OD ratios at boundary inflows are the same for HOVs and SOVs by Assumption 3.2. Thus, the traffic flows at the toll entry can be determined by the toll entry-exit OD demand and the associated entering proportions

$$f_i^{in}(t) = \sum_{j\geq i} f_{ij}^{in}(t) = \sum_{j\geq i} r_{ij}(t)P_{ij}(t) = \sum_{j\geq i} \beta_{ij} r_i(t) P_{ij}(t) = r_i(t) \sum_{j\geq i} \beta_{ij} P_{ij}(t) \; \forall j \geq i, \quad (3.19)$$

where f_{ij}^{in} is the flow enters the HOT lane in cell pair i heading for the toll exit in cell pair j; β_{ij} is the proportion of flow in r_i that may choose the HOT lanes and travel to the toll exit in cell pair j.

When the "control" is entry flows instead of toll rates at toll entries, we can avoid use of the nonlinear and nonconvex Logit functions for the lane-choice probabilities in the control model. This can be achieved by using a proper toll-manipulation model, as discussed below.

3.3 Toll-Manipulation Model

3.3.1 Bijection between tolls and toll entry flows

In a multi-access managed-lane system, we aim to determine time-varying tolls for all the toll entry-exit pairs in order to maintain smooth traffic. Note that if we directly optimize the tolls, the nonlinear and nonconvex lane-choice probability function in (3.17) has to be incorporated into the control model, making it intractable. We have an important observation that while tolls control the proportion of SOVs which chooses the HOT lanes at each toll entry, we can view the resultant toll entry flows $\{f_{ij}^{in}\}$ as "intermediate" input to the system which directly affects the distribution of traffic flow in the system. Notice that as indicated by (3.12) and (3.19) , if we take the toll entry flows $\{f_{ij}^{in}\}$ as the input to the system, then the resulting system model is PWA, so we have relatively efficient tools to deal with the control problem (optimization of the toll entry flows), as will be discussed in detail in Section 3.4). Therefore, if we can translate each f_{ij}^{in} conveniently to a corresponding unique toll $\tau_{ij} > 0$, we can almost as easily solve the original control problem by first solving for the optimal toll entry flows and then translating them to the tolls. In fact, this can be achieved based on the following basic observation.

Lemma 3.4 *$P_{ij}(t)$ is a function of $v_i^G(t)$ and $\tau_{ij}(t)$, $P_{ij}(t) = P_{ij}(v_i^G(t),\ \tau_{ij}(t))$, and given $v_i^G(t)$, P_{ij} is invertible for $\tau_{ij}(t) \geq 0$.*

Proof: It can be verified that given v_i^G, $P_{ij}^L(v_i^G,\ \tau_{ij})$ is continuous and strictly decreasing in $\tau_{ij} \geq 0$. Since P_{ij} is an increasing affine function of P_{ij}^L as defined in (3.18) , it is also a function of v_i^G and τ_{ij}, and in particular continuous and strictly decreasing and hence an invertible in $\tau_{ij} \geq 0$ given v_i^G. ■

Therefore, based on the measured v_i^G and the optimal inflow to the HOT lanes (f_{ij}^{in*}) obtained from the control model (the corresponding optimal entering proportion is P_{ij}^*), the optimal toll τ_{ij}^* can be derived by the inverse of the function $P_{ij}(v_i^G, \cdot)$ (time index t is omitted for clarity):

$$\tau_{ij}^* = P_{ij}^{-1}(v_i^G, P_{ij}^*) = P_{ij}^{-1}(v_i^G, f_{ij}^{in*}/r_{ij}). \tag{3.20}$$

3.3.2 Practical constraints

Lower and upper limits on SOVs' choice probabilities

In practice, there is typically a predefined proper toll cap $\tau_{ij}^{\max} > 0$, so $\tau_{ij} \in [0, \tau_{ij}^{\max}]$. Thus, the choice probability P_{ij}^L cannot be higher than some $P_{ij,up}^L \in (0, 1)$

nor lower than some $P^L_{ij,low} \in (0,1)$. If we simply set $P^L_{ij,up} = P^L_{ij}(v^G_i, 0)$ and $P^L_{ij,low} = P^L_{ij}(v^G_i, \tau^{\max}_{ij})$, then $P^L_{ij,up}$ and $P^L_{ij,low}$ depend on v^G_i by (3.17) , which complicates the control design. However, we can establish proper constant $P^L_{ij,up}$ and $P^L_{ij,low}$ by two practical considerations: 1) τ_{ij} should have decent value when the GP lanes at the toll entry i are congested so that SOVs have a paid option for using the faster HOT lanes (one main purpose of value pricing); 2) τ_{ij} can be $\tau^{\max}_{ij}$ only when the system is heavily congested.

Lemma 3.5 *If $P^L_{ij}(v^G_i, \tau_{ij}) \le P^L_{ij}(v^G_{f,i}, 0)$, then for any $n^G_i > n^G_{c,i}$, we have $\tau_{ij} > 0 \ \forall j \ge i$; and if $P^L_{ij}(v^G_i, \tau_{ij}) \ge P^L_{ij}(0, \tau^{\max}_{ij})$, then for any $n^G_i < n^G_{J,i}$, we have $\tau_{ij} < \tau^{\max}_{ij} \ \forall j \ge i$.*

Proof: We know $\partial P^L_{ij} / \partial v^G_i < 0$ by (3.17), so $P^L_{ij}(v^G_i, 0) > P^L_{ij}(v^G_{f,i}, 0)$ for any $v^G_i < v^G_{f,i}$ (i.e., $n^G_i > n^G_{ci}$) and $P^L_{ij}(v^G_i, \tau^{\max}_{ij}) < P^L_{ij}(0, \tau^{\max}_{ij})$ for any $v^G_i > 0$ (i.e., $n^G_i < n^G_{J,i}$). It follows that $P^L_{ij}(v^G_i, \tau_{ij}) \le P^L_{ij}(v^G_{f,i}, 0)$ is possible only when $\tau_{ij} > 0$ and $P^L_{ij}(v^G_i, \tau_{ij}) \ge P^L_{ij}(0, \tau^{\max}_{ij})$ is possible only when $\tau_{ij} < \tau^{\max}_{ij}$ since $\partial P^L_{ij} / \partial \tau_{ij} < 0$. ■

Therefore, we can choose any proper $P^L_{ij,up}$ and $P^L_{ij,low}$ with $P^L_{ij,up} > P^L_{ij,low}$ and

$$\begin{aligned} P^L_{ij,up} &\le \left\{ 1 + \exp\left[2\alpha_{1i} d_{ij} / (v^G_{f,i} + \bar{v}^G_{ij}) - \alpha_{1i} \sum_{j'=i}^{j} d_{j',j+1} / v^H_{f,j'} + \gamma_{ij} \right] \right\}^{-1} \\ P^L_{ij,low} &\ge \left\{ 1 + \exp\left[2\alpha_{1i} d_{ij} / \bar{v}^G_{ij} - \alpha_{1i} \sum_{j'=i}^{j} d_{j',j+1} / v^H_{f,j'} - \alpha_{2i} \tau^{\max}_{ij} + \gamma_{ij} \right] \right\}^{-1} \end{aligned} \tag{3.21}$$

which are independent of v^G_i and ensure restrictions 1) and 2), thus making the control model practical and convenient to handle. Notice that applying $P^L_{ij,up}$ and $P^L_{ij,low}$ by (3.21) is sufficient for $\tau_{ij}(t)$ to be within $[0, \tau^{\max}_{ij}]$ but still provides much flexibility in flow design (i.e., $P^L_{ij,up}$ is much higher than $P^L_{ij,low}$) since in practice $\tau^{\max}_{ij}$ is sufficiently large and $\bar{v}^G_{ij}$ is notably smaller than $v^G_{f,i}$ (which makes pricing necessary in the first place), e.g., see the numerical example in Section 3.5.

Then by (3.18) the corresponding lower and upper bounds on P_{ij} are

$$P_{ij}^{\min} = (1-\eta_i)P_{ij,low}^{L} + \eta_i,\ P_{ij}^{\max} = (1-\eta_i)P_{ij,up}^{L} + \eta_i. \tag{3.22}$$

The constraint $P_{ij}(t) \in [P_{ij}^{\min}, P_{ij}^{\max}]$ can then be translated to limits on toll entry flows

$$r_{ij}(t)P_{ij}^{\min} = f_{ij}^{in,\min}(t) \le f_{ij}^{in}(t) \le f_{ij}^{in,\max}(t) = r_{ij}(t)P_{ij}^{\max}. \tag{3.23}$$

Constraints for equity considerations and proper use of the ratios $\{\lambda_i\}$

One issue with the OD-based tolling scheme is equity among the potential HOT lane users of different OD pairs [87]. Here we propose the following constraints at each toll entry: the proportions of the flows entering the HOT lanes and going to different downstream exits do not vary significantly (the similarity can be OD-specific), i.e.,

$$(1+\varepsilon_{ij})^{-1}P_{ij} \le P_{i,j+1} \le (1+\varepsilon_{ij})P_{ij} \text{ for some small } \varepsilon_{ij} > 0\ \forall j \in \{i, i+1, ...N-1\}. \tag{3.24}$$

By (3.18) we know that similar P_{ij} means similar P_{ij}^{L}, for all $i \le j \le N$. Hence the intuition behind (3.24) is that the SOVs at the same toll entry heading for different downstream toll exits have similar likelihood to use the HOT lanes. Note that although (3.24) leads to extra restrictions to the tolls τ_{ij} ($i \le j \le N$), it does not limit the level of capacity utilization of the HOT lanes as there is no restriction to the absolute value of any particular P_{ij}. Now we show that (3.24) may also help justifying the use of constant ratio λ_i (see Remark 3.2).

Proposition 3.1 *If (3.24) holds, then $\lambda_j \approx 1 - \beta_{1j}/(\sum_{j' \ge j}\beta_{1j'})$ provided that: (i) $(1 - P_{11}^{\max})r_1 >> (1 - P_{11}^{\min}))r_i$ for all $1 < i \le j$ (if any) or (ii) the OD ratios $\beta_{ij'}$ are close for all $i \le j$ given each $j' \ge j$.*

Proof: Under case (i), we have by definition of λ_j and (3.24)

$$\begin{aligned}\lambda_j &= 1-\frac{\sum_{i\leq j} r_{ij}(1-P_{ij})}{\sum_{i\leq j}\sum_{j'\geq j} r_{ij'}(1-P_{ij'})} \approx 1-\frac{r_{1j}(1-P_{1j})}{\sum_{j'\geq j} r_{1j'}(1-P_{ij'})}\\ &= 1-\frac{r_1\beta_{1j}(1-P_{1j})}{\sum_{j'\geq j} r_1\beta_{1j'}(1-P_{ij'})} \approx 1-\frac{\beta_{1j}}{\sum_{j'\geq j}\beta_{1j'}}\end{aligned}$$

and under case (ii), we have by (3.24)

$$\begin{aligned}\lambda_j &= 1-\frac{\sum_{i\leq j} r_i\beta_{ij}(1-P_{ij})}{\sum_{i\leq j}\sum_{j'\geq j} r_i\beta_{ij'}(1-P_{ij'})} \approx 1-\frac{\beta_{1j}\sum_{i\leq j} r_i(1-P_{ij})}{\sum_{i\leq j} r_i\sum_{j'\geq j}\beta_{1j'}(1-P_{ij'})}\\ &\approx 1-\frac{\beta_{1j}\sum_{i\leq j} r_i(1-P_{ij})}{\sum_{i\leq j} r_i(1-P_{ij})\sum_{j'\geq j}\beta_{1j'}} \approx 1-\frac{\beta_{1j}}{\sum_{j'\geq j}\beta_{1j'}}\end{aligned}$$

■

Note that r_1 dominates r_i ($i > 1$) by our modeling choice (see Remark 3.1) and $p_{ij}^{\max}$ should be not too close to 1 nor $p_{ij}^{\max}$ to 0 by use of (3.21) in setting $p^L_{ij,up}$ and $p^L_{ij,low}$; and if two toll entries i and i' are close (which is the model setting of our focus) the underlying OD ratios $\{\beta_{ij}\}$ and $\{\beta_{i'j}\}$ to downstream exits could be similar. Hence imposing (3.24) also ensures that $\lambda_j \approx 1-\beta_{1j}/(\sum_{j'\geq j}\beta_{1j'})$ is a good approximation by Proposition 3.1. Notice that here the first cell pair actually represents the location of a major inflow to the modeling segment, hence this approach can be used for estimating λ_j's for a long system with several modeling segments (see Remark 3.1).

Relative relationships among tolls to different exits

Another practical requirement in the OD-based tolling [87] is that at each toll entry the toll for a more distant toll exit cannot be lower than that for a closer toll exit. We show below that this restriction has a good property (Proposition 3.2) that admits a convenient sufficient condition.

Lemma 3.6 *For any j and j' with $i \leq j < j' \leq N$, the perceived travel time satisfies $t_{ij} < t_{ij'}$.*

Proof: First we let $\bar{t}_{ij}^G$ be the historical average travel time users experienced on distance d_{ij} via GP lanes, so $\bar{v}_{ij}^G = d_{ij}/\bar{t}_{ij}^G$. Showing $t_{ij} < t_{ij'}$ is equivalent to showing $d_{ij}/(v_i^G + \bar{v}_{ij}^G) < d_{ij'}/(v_i^G + \bar{v}_{ij'}^G)$, which is equivalent to $d_{ij}/(v_i^G + d_{ij}/\bar{t}_{ij}^G) < d_{ij'}/(v_i^G + d_{ij'}/\bar{t}_{ij'}^G)$, and this is true because

$$\left(\frac{d_{ij}}{v_i^G + d_{ij}/\bar{t}_{ij}^G}\right)^{-1} = \frac{v_i^G}{d_{ij}} + \frac{1}{\bar{t}_{ij}^G} > \frac{v_i^G}{d_{ij'}} + \frac{1}{\bar{t}_{ij'}^G} = \left(\frac{d_{ij'}}{v_i^G + d_{ij}/\bar{t}_{ij'}^G}\right)^{-1}$$

since $d_{ij} < d_{ij'}$ and $\bar{t}_{ij}^G < \bar{t}_{ij'}^G$. ■

Proposition 3.2 *Let $S_i = \{(f_{ij}^{in}, f_{ij'}^{in}) : \tau_{ij} \leq \tau_{ij'}$ for any $j' > j] \geq i\}$, then given r_i and v_i^G, S_i is convex.*

Proof: By (3.17) we have that for different toll exits j and j' with $j' > j \geq i$ (time index t omitted),

$$\ln\left((P_{ij}^L)^{-1} - 1\right) = \alpha_{1i} t_{ij} + \gamma_{ij} - \alpha_{2i}\tau_{ij}, \; \ln\left((P_{ij'}^L)^{-1} - 1\right) = \alpha_{1i} t_{ij'} + \gamma_{ij'} - \alpha_{2i}\tau_{ij'}. \quad (3.25)$$

Given r_{ij}, P_{ij}^L is a positive affine function of P_{ij} by (3.18), and hence a positive affine function of f_{ij}^{in}. Then by (3.25), $\tau_{ij} \leq \tau_{ij'}$ implies that implies that

$$\alpha_{2i}\tau_{ij} = \alpha_{1i} t_{ij} + \gamma_{ij} - \ln\left((P_{ij}^L)^{-1} - 1\right) \geq \alpha_{1i} t_{ij'} + \gamma_{ij'} - \ln\left((P_{ij'}^L)^{-1} - 1\right) = \alpha_{2i}\tau_{ij'},$$

where the inequality is due to $\tau_{ij} \leq \tau_{ij'}$ and $\alpha_{2i} < 0$. By rearranging terms, we obtain

$$\ln\left[\left((P_{ij'}^L)^{-1} - 1\right) / \left((P_{ij}^L)^{-1} - 1\right)\right] \geq \alpha_{1i}(t_{ij'} - t_{ij}) + (\gamma_{ij'} - \gamma_{ij}).$$

This implies that (let $\omega = \exp[\alpha_{1i}(t_{ij'} - t_{ij}) + (\gamma_{ij'} - \gamma_{ij})]$)

$$P_{ij'}^L \leq f(P_{ij}^L) = P_{ij}^L\left(\omega + (1-\omega)P_{ij}^L\right)^{-1}, \quad (3.26)$$

where we define the rightmost term as a function of P^L_{ij}, $f(P_ij^L)$, whose second derivative is

$$\frac{\partial^2 f}{\partial (P^L_{ij})^2} = \partial\left(\frac{\omega}{((1-\omega)P^L_{ij}+\omega)^2}\right)/\partial P^L_{ij} = -\frac{2\omega(1-\omega)}{((1-\omega)P^L_{ij}+\omega)^3}. \tag{3.27}$$

Since $\alpha_{1i} \leq 0$, $t_{ij} < t_{ij'}$ (by Lemma 3.6) and $\gamma_{ij'} \leq \gamma_{ij} \leq 0$ (by Assumption 3.3), we have $0 \leq \omega \leq 1$, thus (3.27) implies that $\partial^2 f/\partial (P^L_{ij})^2 \leq 0$. Thus, f is concave, so (3.26) is convex in $(P^L_{ij}, P^L_{ij'})$ and thus also convex in $(f^{in}_{ij}, f^{in}_{ij'})$ since given r_i, P^L_{ij} is a positive affine function of f^{in}_{ij} for all $j \geq i$, so S_i is convex. ■

Therefore, to have $\tau_{ij} \leq \tau_{ij'}$ for any $j' > j \geq i$, only the following constraints are needed for each i:

$$\frac{f^{in}_{ij} - r_{ij}\eta_i}{(1-\eta_i)r_{ij}}\left(\omega + \frac{(1-\omega)(f^{in}_{i,j-1} - \eta_i r_{i,j-1})}{(1-\eta_i)r_{i,j-1}}\right) \leq \frac{f^{in}_{i,j-1} - r_{i,j-1}\eta_i}{(1-\eta_i)r_{i,j-1}} \;\forall j > i. \tag{3.28}$$

Based on the convexity of (3.28), we can derive a linear constraint that ensures (3.28) for control design. One natural option is to connect the two corner points of the feasible region for $(P_{i,j-1}, P_{ij})$ defined by (3.22): $(P^{\min}_{i,j-1}, P^{\min}_{ij})$, and $(P^{\max}_{i,j-1}, P^{\max}_{ij})$, for $j > i$, and require $P_{ij} \leq (P_{i,j-1} - P^{\min}_{i,j-1})(P^{\max}_{ij} - P^{\min}_{ij})/(P^{\max}_{i,j-1} - P^{\min}_{i,j-1}) + P^{\min}_{ij}$. These restrictions amount to a set of linear constraints on $\{f^{in}_{ij}\}$ for each toll entry i:

$$\frac{f^{in}_{ij}}{r_{ij}} \leq \frac{P^{\max}_{ij} - P^{\min}_{ij}}{P^{\max}_{i,j-1} - P^{\min}_{i,j-1}}\left(\frac{f^{in}_{i,j-1}}{r_{i,j-1}} - P^{\min}_{i,j-1}\right) + P^{\min}_{ij} \;\forall j > i. \tag{3.29}$$

Notice that the coefficients in (3.28) depend on v^G_i via variable ω. but (3.29) is independent of v^G_i, a nice property for control design. Figure 3.2 depicts the two constraints (3.28) and (3.29) in terms of P_{ij} for an illustrative example (toll entry i and toll exits j and j' with $i \leq j \leq j'$). We can see that the feasible region for the linear constraint (3.29) lies within and reasonably approximates the feasible region for the convex constraints (3.28) for a wide range of different v^G_i's. We will impose (3.29) in the control model.

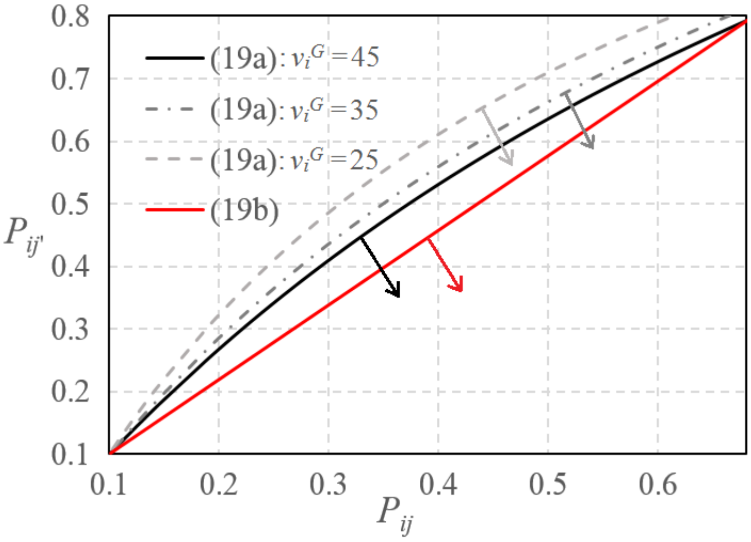


Figure 3.2: Toll constraints in terms of P_{ij} and $P_{ij'}$ (arrows indicate the feasible regions; data used: $d_{ij} = 1$, $d_{ij'} = 2$, $v^H_{f,i} = v^G_{f,i} = 60$, $\bar{v}^G_{ij} = 30$, $\alpha_{1i} = 20$, $\gamma_{ij} = -0.5d_{ij}$, $\tau^{\max}_{ij} = \tau^{\max}_{ij'} = \infty$, $\eta_i = 0.1$).

One key observation is that our control model is flexible enough to incorporate such constraints as (3.23), (3.24) and (3.29), which will be explained in Section 3.4.3 .

Remark 3.4 *It can be verified that at each toll entry the fairness condition (3.24) is compatible with the condition of no less toll for a more distant toll exit. To see this, we note that for any j and j' with $i \leq j < j' \leq N$, if $P_{ij'} = P_{ij}$, then (3.17) implies that $\alpha_{1i}t_{ij} - \alpha_{2i}\tau_{ij} + \gamma_{ij} = \alpha_{1i}t_{ij'} - \alpha_{2i}\tau_{ij'} + \gamma_{ij'}$. Since $d_{ij} < d_{ij'}$, $\alpha_{2i} < 0$, $\alpha_{1i} \leq 0$, $\gamma_{ij'} \leq \gamma_{ij}$ (by Assumption 3.3 since $d_{ij} < d_{ij'}$), $t_{ij} < t_{ij'}$ (by Lemma 3.6), we deduce that* $\tau_{ij} - \tau_{ij'} = [\alpha_{1i}(t_{ij} - t_{ij'}) + \gamma_{ij} - \gamma_{ij'}]/\alpha_{2i} \leq 0$.

Remark 3.5 *For easy of notation, we have assumed that all cell pairs have toll entry and exit, the derivations can easily be modified so that only the cells that actually have these components appear in the equations. We will keep use this convenient notation,* *and without loss of generality, let $V = (N+1)N/2$ be the maximum possible number of toll entry/exit pairs, typically it is much smaller.*

3.4 The Control Model and Solution Method

3.4.1 System predictive model

Let the system state be $x = [(n^H)^{\mathrm{T}}, (n^G)^{\mathrm{T}}, l]^{\mathrm{T}} \in \mathbb{R}^{2N+3}$ and the controlled input be $u = \{f_{ij}^{in}\} \in \mathbb{R}^V$, then based on (3.12), (3.14) and (3.15) as well as the linear relationship $f_i^{in} = \sum_{j \geq i} f_{ij}^{in}$, we can describe the entire system by the following relations:

$$x(t+1) = Ax(t) + B_1 u(t) + B_2 w(t) + B_3 \delta(t) + B_4 s(t) + \epsilon(t), \tag{3.30}$$

$$C_1 \delta(t) + C_2 s(t) \leq C_3 u(t) + C_4 w(t) + C_5 x(t) + C_6, \tag{3.31}$$

where $\delta = \{\delta_k,\ k = 1, ..., M\} \in \{0,1\}^M$, $s = \{s_k,\ k = 1, ..., M\} \in \mathbb{R}^M$, with each binary auxiliary variable δ_k and each continuous auxiliary variable s_k defined earlier; A, $B_{1\sim 4}$, $C_{1\sim 6}$ are constant matrices of suitable dimensions; ϵ is the random noise term that represents modeling error (e.g., modeling the real traffic flow by CTM and use of constant λ_i) and demand forecast error and implementation error (use of Logit lane-choice probability functions), we assume that $\epsilon(t)$ has zero mean, is bounded and independent across t. Given the current state $x(t)$ and input $u(t)$, the evolution (3.31) is determined by a feasible value of $\delta(t)$ and $s(t)$ to (3.31).

3.4.2 Control model formulation

Two major advantages of MPC are: 1) explicit consideration of and computation with state and input constraints which would be very hard to accomplish in any other way [17, 22]; 2) effective disturbance rejection by adjusting the action based on current state measurement and a certain length of system prediction. At each control stage t, given the observation $x_t = x(t)$, the planned control sequence $u_t, u_{t+1}, ..., u_{t+P1}$, and the exogenous input forecast $w_{t+1}, ..., w_{t+P-1}$, we can predict the future states over the prediction horizon P as $x_{t+1}, ..., x_{t+P}$. We use subscript $t+p$ to indicate a quantity predicted or to be computed p-step into the future based on the current observation x_t. At each stage $t \geq 0$ we compute the controls $u_t, u_{t+1}, ..., u_{t+P1}$ by solving the following optimization problem:

$$\min_{\substack{u_t,...,u_{t+P-1} \in \mathbb{R}^V \\ \delta_t,...,\delta_{t+P-1} \in \{0,1\}^M}} \quad J_0(t) = \sum_{p=1}^{P} b^{\mathrm{T}}(x_{t+p} - B_0 n_{t+p}^{HOV}) + o^{\mathrm{T}} n_{t+p}^{HOV} + \rho \| u_{t+p} - u_{t+p-1} \|_1 \tag{3.32}$$

$$\text{s.t.} \quad x_t = x(t), \tag{3.33}$$

$$x_{t+p+1} = A x_{t+p} + B_1 u_{t+p} + B_2 w_{t+p} + B_3 \delta_{t+p} + B_4 s_{t+p}, \quad p = 0, ..., P-1, \tag{3.34}$$

$$C_1 \delta_{t+p} + C_2 s_{t+p} \leq C_3 u_{t+p} + C_4 w_{t+p} + C_5 x_{t+p} + C_6, \quad p = 0, ..., P-1, \tag{3.35}$$

$$x^{\min} \leq x_{t+p} \leq x^{\max}, \quad p = 1, ..., P, \tag{3.36}$$

$$\Delta u^{\min} \leq u_{t+p} - u_{t+p-1} \leq \Delta u^{\max}, \quad p = 0, ..., P-1, \tag{3.37}$$

$$u_{t+p}^{\min}(w_{t+p}) \leq u_{t+p} \leq u_{t+p}^{\max}(w_{t+p}), \quad p = 0, ..., P-1, \tag{3.38}$$

$$F(u_{t+p}, w_{t+p}) \leq 0 \quad p = 0, ..., P-1. \tag{3.39}$$

The coefficients of the objective function $J_0(t)$ in (3.32) are $B_0 = [I, \mathbf{0}]^{\mathrm{T}}$, $b = [\mathbf{1}^{\mathrm{T}}, 0, \mathbf{1}^{\mathrm{T}}, \mathbf{0}, 1]^{\mathrm{T}}$, here I is a N-by-N identity matrix, $\mathbf{0}$ is a $(N+3)$-by-$(N+3)$ zero matrix, $n^{HOV} = [n_1^{HOV}, ..., n_N^{HOV}]^{\mathrm{T}}$ and $o = [o_1, ..., o_N]^{\mathrm{T}}$, respectively, contain the number of HOVs and the average occupancy of HOVs on each HOT lane cell.

Thus by Assumption 3.2, the first two terms in the summation in $J_0(t)$ are the total number of travelers in the freeway segment at time $t+p$. Hence in (3.32) we want to minimize the weighted sum of two terms over the prediction horizon: the total person travel time and the effort involved in control input changes, with $\rho > 0$ represents the weight of the control smoothing term relative to the total person delay term. The constraints are described below.

(3.33)-(3.35) are system dynamics constraints based on (3.31) and (3.31), where we replace the noise $\epsilon(t+p)$ with its expectation $\mathbf{0}$, this is a commonly used approach in practice for its convenience and good empirical performance in MPC [17].

(3.36) is the state constraint. We restrict the traffic density to be at most the critical value for HOT lanes, so $x^{\min} = \mathbf{0}$, $x^{\max} = [n^H_{c,1}, ..., n^H_{c,N}, \bar{n}^{\mathrm{T}}]^{\mathrm{T}}$.

(3.37) applies the limits $\Delta u^{\min}$ and $\Delta u^{\max}$ on delta changes in the input [120], which can determined based on traffic safety and stability needs.

(3.38) are the limits (3.23) on the input u. Note that we write the bounds on u_{t+p} as functions of w_{t+p}, because they depend on r_{ij}.

(3.39) encodes constraints (3.24) and (3.29) in Propositions 3.1 and 3.2. F is a vector-valued function with suitable dimension.

Although all P controls are computed in the optimization model, only the first one, u_t, is implemented while the others are discarded. The same process is repeated in the next stage, once it is revealed.

3.4.3 Properties and solution method of the control model

We first reformulate problem (3.32)-(3.39).

Proposition 3.3 *Problem (3.32)-(3.39) is equivalent to the following (where $\phi_{t+p} = [\phi_{t+p,1}, ..., \phi_{t+p,V}]^{\mathrm{T}}$):*

$$\min_{\substack{u_t,...,u_{t+P-1},\phi_t,...,\phi_{t+P-1}\in\mathbb{R}^V \\ \delta_t,...,\delta_{t+P-1}\in\{0,1\}^M}} J(t) = \sum_{p=1}^{P} b^{\mathrm{T}} x_{t+p} + \rho \mathbf{1}^{\mathrm{T}} \phi_{t+p} \tag{3.40}$$

$$\text{s.t. } (3.33)-(3.39), -\phi_{t+p,l} \leq u_{t+p,l} - u_{t+p-1,l} \leq \phi_{t+p,l}, p = 1, ..., P-1, l = 1, ..., V \tag{3.41}$$

Proof: Since the HOT lanes are maintained at free-flow condition, by Assumptions 3.2 and 3.3, the number of HOVs in the HOT lane cells can be predicted by (where A^H is defined in (3.14))

$$n^{HOV}(t+1) = A^H n^{HOV}(t) + \mathrm{diag}(\eta_1, ..., \eta_N) r(t).$$

It follows that the terms which involve n^{HOV} can be dropped from the objective function $J^0(t)$ in (3.32) without affecting its optimal solution.

Also note that $|u_{t+p,l} - u_{t+p-1,l}| \leq \phi_{t+p,l}$ is equivalent to $-\phi_{t+p,l} \leq u_{t+p,l} - u_{t+p-1,l} \leq \phi_{t+p,l}$ for each $p = 0, ..., P-1$ and $l = 1, ..., V$, where $u_{t+p,l}$ is the l^{th} entry of u_{t+p}. In addition, at the optimal solution to problem (3.40)-(3.41) we must have $|u_{t+p,l} - u_{t+p-1,l}| = \phi_{t+p,l}$, since otherwise the objective function $J(t)$ in (3.40) can be improved by reducing $\phi_{t+p,l}$. This completes the proof. ■

Although more variables are involved in problem (3.40)-(3.41), it is easier to handle than problem (3.32)-(3.39) as occupancy data (which is usually hard to obtain) is not needed in computing the optimal control while the resulting optimal solution also minimizes the original objective $J_0(t)$ in (3.32) and the new objective $J(t)$ in (3.40) is clearly linear in x_{t+p} and ϕ_{t+p-1}, $p = 1, ..., P$.

Now we examine the structure of the constraints in (3.41). Clearly, the constraints in (3.33)-(3.37) are all linear in the state x_{t+p}, $p = 1, ..., P$, the control u_{t+p}, $p = 0, ..., P-1$, and the auxiliary binary variables δ_{t+p}, $p = 0, ..., P-1$. The predicted exogenous input w is not involved in these constraints. For the rest of the constraints, we claim the following.

Proposition 3.4 *The constraints in (3.38) (which encode (3.23)) and the constraints in (3.39) (which encode (3.24) and (3.29) are all linear in both u_{t+p} and w_{t+p} for $p = 0, ..., P-1$.*

Proof: Under Assumption 3.2 and (3.22), we know that $P_{ij}^{\min}$, $P_{ij}^{\max}$ are constant within the problem horizon, hence by (3.23) we know that $f_{ij,t+p}^{in,\min} \leq f_{ij,t+p}^{in} \leq f_{ij,t+p}^{in,\max}$ is equivalent to

$$f_{ij,t+p}^{in} - \beta_{ij} r_{i,t+p} P_{ij}^{\max} \leq 0, \; -f_{ij,t+p}^{in} + \beta_{ij} r_{i,t+p} P_{ij}^{\min} \leq 0.$$

Hence the constraints in (3.38) are linear in both u_{t+p} and w_{t+p} for $p = 0, ..., P-1$. Assumption 3.2 implies that r_{ij} and $r_{i,j+1}$ are fixed proportions of the total bypass inflow r_i within the problem horizon, thus for each $p = 0, ..., P-1$, (3.24) is equivalent to

$$\begin{aligned} & (1+\varepsilon_{ij})^{-1} \frac{f_{ij,t+p}^{in}}{\beta_{ij} r_{i,t+p}} \leq \frac{f_{i,j+1,t+p}^{in}}{\beta_{i,j+1} r_{i,t+p}} \leq (1+\varepsilon_{ij}) \frac{f_{ij,t+p}^{in}}{\beta_{ij} r_{i,t+p}} \; \forall i \leq j < N \\ \Leftrightarrow \quad & \beta_{ij} f_{i,j+1,t+p}^{in} - (1+\varepsilon_{ij}) \beta_{i,j+1} f_{i,j,t+p}^{in} \leq 0, \\ & (1+\varepsilon_{ij})^{-1} \beta_{i,j+1} f_{ij,t+p}^{in} - \beta_{ij} f_{i,j+1,t+p}^{in} \leq 0 \; \forall i \leq j < N. \end{aligned}$$

Similarly, for each i, each $j > i$ and each $p = 0, ..., P-1$, (3.29) is equivalent to

$$\begin{aligned} & \frac{f_{ij,t+p}^{in}}{\beta_{ij} r_{i,t+p}} - \frac{P_{ij}^{\max} - P_{ij}^{\min}}{P_{i,j-1}^{\max} - P_{i,j-1}^{\min}} \frac{f_{i,j-1,t+p}^{in}}{\beta_{i,j-1} r_{i,t+p}} - \frac{P_{i,j-1}^{\max} P_{ij}^{\min} - P_{ij}^{\max} P_{i,j-1}^{\min}}{P_{i,j-1}^{\max} - P_{i,j-1}^{\min}} \leq 0 \\ \Leftrightarrow \quad & \beta_{i,j-1}(P_{i,j-1}^{\max} - P_{i,j-1}^{\min}) f_{ij,t+p}^{in} - \beta_{ij}(P_{ij}^{\max} - P_{ij}^{\min}) f_{i,j-1,t+p}^{in} \\ & -\beta_{i,j-1}(P_{i,j-1}^{\max} P_{ij}^{\min} - P_{ij}^{\max}) P_{i,j-1}^{\min}) r_{i,t+p} \leq 0. \end{aligned}$$

Thus F is affine, so the constraints in (3.39) are all linear in u_{t+p} and w_{t+p} for $p = 0, ..., P-1$. ■

We define the decision vector $z_t = [x_{t+1}^{\mathrm{T}}, ..., x_{t+P}^{\mathrm{T}}, u_t^{\mathrm{T}}, ..., u_{t+P-1}^{\mathrm{T}}, \pi_0^{\mathrm{T}}, ..., \phi_{P-1}^{\mathrm{T}}, \delta_t^{\mathrm{T}}, ..., \delta_{t+P-1}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{P(2N+3+2V)} \times \{0,1\}^{PM}$, and the vector of parameters $y_t = [x(t), w_t^{\mathrm{T}}, ..., w_{t+P-1}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2N+3+P(N+1)}$ that contains the current state and the predicted boundary inflows within the prediction horizon. Then by Proposition 3.4.3, problem (3.40)-(3.41) can be cast as a multi-parametric mixed integer linear programing (mp-MILP) problem

$$J^*(y_t) = \min_{z_t}\{J(z_t, y_t) = f^{\mathrm{T}} z_t\} \tag{3.42}$$
$$\text{s.t. } Gz_t \leq S y_t + g,$$

where $f = [b^{\mathrm{T}}, ..., b^{\mathrm{T}}, \mathbf{1}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{P(2N+3+2V+M)}$, and $G \in \mathbb{R}^{Q \times P(2N+3+2V+M)}$, $S \in \mathbb{R}^{Q \times (2N+3+P(N+1))}$, $g \in R^Q$ are constant coefficients which can be constructed based on the data in problem (3.40)-(3.41). Here Q is the total number of inequalities needed to express the constrains in (3.41). Let $\Theta \subseteq \mathbb{R}^{2N+3+P(N+1)}$ be the polytope formed by the upper and lower bounds of the entries in y and denote $\Theta^* \subseteq \Theta$ the region of parameters $y \in \Theta$ such that problem (3.40)-(3.41) is feasible. Notice that by Assumption 3.2 we have $\Theta^* \neq \emptyset$ in the first place. For any given $\bar{y}_t \in \Theta^*$, $J^*(\bar{y}_t)$ denotes the minimum value of $J(y_t)$ for $y_t = \bar{y}_t$. The value function $J^* : \Theta* \rightarrow \mathbb{R}$ denotes the function which expresses the dependence on y of the minimum value of the objective function over $\Theta*$. The set-valued function $Z^* : \Theta^* \rightarrow 2^{\mathbb{R}^{P(2N+3+2V)}} \times 2^{\{0,1\}^{PM}}$ describes for any fixed $y_t \in \Theta^*$ the set of optimizers $z^*(y_t)$ corresponding to $J^*(y_t)$.

If we can determine the region Θ^* of feasible parameters y and find the expressions of value function $J^*(y_t)$ and an optimizer function $z_t^*(y_t) \in Z^*(y_t)$, then computing the optimal control at each step t amounts to evaluating an analytical form. Now we claim that this actually can be achieved due to the property of mp-MILP.

Theorem 3.2 *Consider problem (3.42). The set Θ^* is the union of a finite number of mutually disjoint polyhedra and the value function J^* is PWA on these polyhedra. In addition, it is always possible to define a PWA optimizer function $Z^*(y_t)$ on these polyhedra.*

Proof: Given any set of fixed binary variables $\delta = [\delta_t^{\mathrm{T}}, ..., \delta_{t+P-1}^{\mathrm{T}}]^{\mathrm{T}}$, problem (3.42) becomes a multi-parametric linear programing (mp-LP) problem, whose optimal value function $J^*(y_t, \delta)$ is known to be PWA on a set of polydera $\{P_i(\delta)\}$ (a partition of Θ^*) and it is always possible to define a continuous and PWA optimizer function $Z^*(y_t, \delta)$ on Θ^* [22]. Superimposing all the polydera $\{P_i(\delta)\}$ for all $\delta \in \{0, 1\}^{PM}$ with $\sum_{k=1,...,M} \delta_{tk} = 1$, $t = 1, ..., P$, we obtain a more refined polydera partition $\{P_i\}$ of $\Theta *$ since the intersection of any two polyhera is polyhedral. Hence for any P_i and any $y \in P_i$, $J^*(y)$ is the minimum over a finite set of affine functions in $\mathbb{R}^{2N+3+P(N+1)} \rightarrow \mathbb{R}$, each of which corresponds to one distinct δ, so $J^*(y)$ is affine on P_i. Figure 3.3 shows a simple example of PWA value function $J^*(y)$ when parameter y is a scalar and the number of feasible binary vectors δ is 2. Now we write $Z^*(y) = (Z_c^*(y), Z_d^*(y))$ with the continuous part $Z_c^*(y) : \Theta^* \rightarrow \mathbb{R}^{P(2N+3+2V)}$ and the discrete part $Z_d^*(y) : \Theta^* \rightarrow \{0, 1\}^{PM}$. Using similar argument, we deduce that it is always possible to define a PWA $Z_c^*(y)$ and a piecewise constant $Z_d^*(y)$ for all $y \in Theta^*$, hence a PWA optimizer function $Z^*(y)$ over Θ^* (since piecewise constant is a special case of PWA). ■

Therefore, problem (3.40)-(3.41) admits an explicit feedback control law, as summarized below.

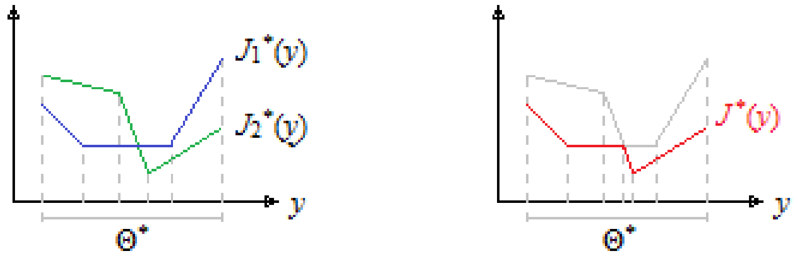


Figure 3.3: Example of the PWA value function when $\Theta \subseteq \mathbb{R}$, $P = 1$, and $M = 2$ ($J_i^*(y)$ is the optimal value function under $\delta = \delta_i$, $i = 1, 2$).

Corollary 3.1 *There exists an optimal control sequences $u^* = [u_t^{*\mathrm{T}}, ..., u_{t+p-1}^{*\mathrm{T}}]^{\mathrm{T}}$ specified by (3.43) and achieves the minimum value for problem (3.40)-(3.41):*

$$u_k^*(x(t), w_t, ..., w_{t+P-1}) = H_k[x(t), w_t, ..., w_{t+P-1}]^{\mathrm{T}} + h_k,$$

$$\text{if}[x(t), w_t, ..., w_{t+P-1}]^{\mathrm{T}}] \in \Theta_k^*, \;\; k = 1, ..., K, \tag{3.43}$$

where Θ_K^*, $k = 1, ..., K$, form a polyhedral partition of the set Θ^*, and H_k, h_k are constant coefficients with suitable dimensions.

The size of the decision vector z in (3.42) is linear in M, so exponentially in N (Lemma 3.2). Moreover, both the size of z and number of constraints in (3.42) grows linearly in P. Thus, for larger system and longer prediction horizon, solving (3.42) for a given parameter vector y can be costly, in which case an explicit solution (3.43) obtained from off-line computation is desirable. An geometric algorithm [37] can be used for solving (3.42) off-line, which is based on a recursion between the solution of an mp-LP subproblem and an MILP subproblem. However, when N and P are relatively small, an on-line controller can also be implemented conveniently. This is because small-to-intermediate sized MILP can be solved efficiently using methods such as the branch and bound algorithm that avoids complete enumeration of the exponentially many combinations of the binary variables [22].

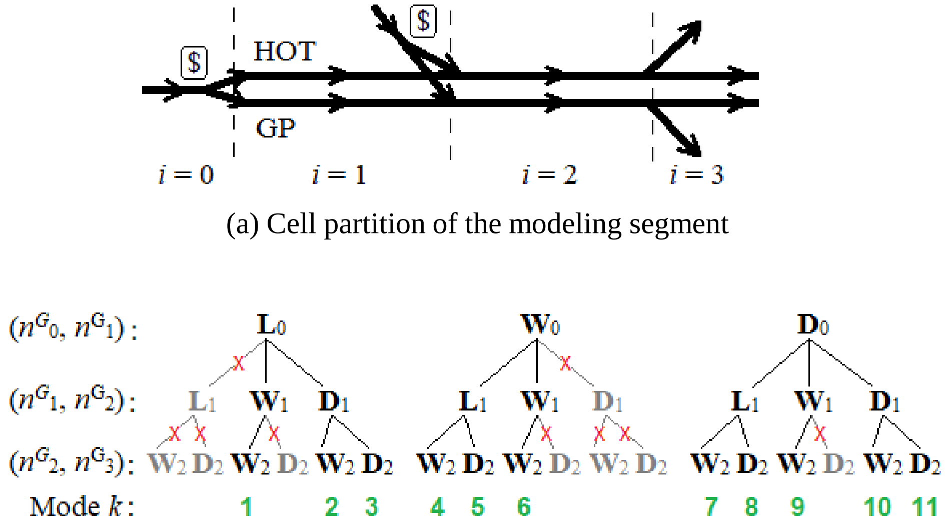
The model parameters such as WTP, λ_i, β_{ij} and η_i, can be updated once a while based on new data using approaches such as the one proposed in [134], then a new problem (3.42) with these updated parameters can be defined and solved.

3.5 Numerical Example

3.5.1 Data and simulation setup

In this section, we conduct a simulation study of the proposed tolling algorithm for a numerical example. In this example, we have a 3-mile long freeway segment with two GP lanes and two HOT lanes, which included two explicit cell pairs for traffic modeling (so $N = 2$), and density on the last mile of the freeway segment represents the downstream boundary condition, thus $n_3^G(t)$ is given. There are two toll entries: one at the start of the first cell pair and the other upstream the ramp access to the second cell pair. Two toll exits are considered, one is located at the off-ramp connected to the second cell pair (exit via either HOT lanes or GP lanes) and the other one is the exit of the modeling segment. Figure 3.4(a) shows the studied freeway segment, we use this basic example to focus on verifying the effectiveness of the proposed control model. We assume that the freeway cells have homogeneous parameters v_f, w, n_J, n_c and q_M (for both HOT and GP lanes), e.g., the free-flow speed is 60 miles per hour in the system. Each modeling step is $t = 0.5$ (min).

By Remark 3.3, we deduce that the modes $m_1 = 1 = (L_0, L_1)$ for cell pair $(0, 1)$, $m_1 = 6 = (W_0, D_1)$ and $m_2 = 6 = (W_1, D_2)$ for cell pair $(1, 2)$ (see Table 3.2) can be removed, thus we only have to consider $M = 11$ mode vectors m_k $(k = 1, ..., 11)$ as shown in Figure 3.4(b).



(a) Cell partition of the modeling segment

(a) Necessary GP lane traffic modes in the PWA system representation

Figure 3.4: Cell partition of the modeling segment and the possible mode vectors. (The gray branches with red crossings are "pruned" from the mode tree).

We assume that in making lane choice decisions the factors considered by SOVs other than travel time and toll is the travel time variation, which is higher in GP lanes than HOT lanes and is proportional to distance traveled [86], so we can write $\gamma_{ij} = \gamma_i d_{ij}$ with $\gamma_i \leq 0$. We set $\alpha_{2i} = 1$, as utility is generally measured in monetary units, then the VOT equals $-\alpha_{1i}$, which we assumed to be 20\$/hr. The willingness to pay for traveling the distance d in the HOT lanes starting at cell pair i that generates one unit of time savings on average (i.e., $d^* = 1/[(\bar{v}_i^G)^{-1} - (v_i^H)^{-1})]$, WTP$_i$, is estimated to be 50\$/h according to [24]. We set $\bar{v}_i^G = 30$ mile/h, so we infer $\gamma_i = -(\text{WTP}_i + \alpha_{1i})/d^* = -0.5$. The maximum toll is $\tau_{ij}^{\max} = 8\$$. The bounds on P_{ij} can be determined by (3.22) based on η_i, $P^L_{ij,low}$ and $P^L_{ij,up}$ (by picking the tightest values with three decimal places according to (3.21). Here we have $P_{1,2}^{\max} = PP_{2,3}^{\max}$ and $P_{1,2}^{\min} = PP_{2,3}^{min}$ since $d_{1,2} = d_{2,3} = 2$ and $\eta_1 = \eta_2 = 0.1$. Also note that since $\beta_{1,2} = \beta_{2,2} = 0.25$ and $\beta_{1,3} = \beta_{2,3} = 0.75$, by Proposition 3.1 we know that the constraint (3.24) can ensure very good accuracy by using $\lambda_2 = 0.75$. These model parameters are summarized in Table 3.3.

Table 3.3: Parameters used in the numerical example

$L_1 = L_2 = L_3$	$\eta_1 = \eta_2$	$\beta_{1,2} = \beta_{2,2}$	$\beta_{1,3} = \beta_{2,3}$	λ_1	λ_2	q_M
1	0.1	0.25	0.75	1	0.75	30
$v,\ w$	$n_c,\ n_J$	$\alpha_{11} = \alpha_{12}$	$\alpha_{21} = \alpha_{22}$	$\gamma_1 = \gamma_2$	$\tau_{ij}^{\max}$	ε_{ij}
0.5, 0.375	60, 140	-20	-1	-0.5	8	0.2
$P_{1,2}^{\max} = P_{2,3}^{\max}$	$P_{1,3}^{\max}$	$P_{2,2}^{\max}$	$P_{1,2}^{\min} = P_{2,3}^{\min}$	$P_{1,3}^{\min}$	$P_{2,2}^{\min}$	
0.795	0.875	0.683	0.107	0.107	0.127	0.102

Based on $P_{ij}^{\min}$, $P_{ij}^{\max}$, ε_{ij}, and OD ratios β_{ij}, we can compute the coefficients of the linear constraints in (3.41) for the input $\{f_{ij}^{in}\}$ and $\{r_i\}$. The delta change limits are $\Delta u^{\min} = 5 \times \mathbf{1}_{N\times 1}$ and $\Delta u^{\max} = 5 \times \mathbf{1}_{N\times 1}$. The weight is $\rho = 0.1$. The choice of prediction horizon P will be discussed later.

We validate the effectiveness of the proposed hybrid MPC model using two scenarios: 1) the system starts with uncongested traffic and then the demand upstream the main toll entry increases significantly together with formation of down- stream traffic jam; 2) the system starts with prevalent GP lane congestion (due to downstream traffic jam) and then the downstream congestion dissipates. Table 3.4 describes the demand profile and the initial condition for each case. We assume that the boundary inflow or density conditions change quickly and then stay constant again. This is a typical way of evaluating the effectiveness of the traffic control approach, as was adopted in many studies, e.g., [12, 55, 83, 120].

In addition, these basic representative scenarios are useful for understanding the mechanism behind our dynamic tolling strategy in response to a certain change. The real traffic conditions may consist of a mixture or a sequence of such changes, hence it is important to see how our controller deal with a basic change in demand or boundary conditions.

Table 3.4: Data of the two traffic scenarios

Scenario	State vector $x = [(n^H)^{\mathrm{T}}, (n^G)^{\mathrm{T}}, l]^{\mathrm{T}} = [n_1^H, n_2^H, n_0^G, n_1^G, n_2^G, n_3^G, l]^{\mathrm{T}}$
#1	$x(0) = [20, 30, 10^a, 20, 30, 117.5^b, 0]^{\mathrm{T}}$
	Exogenous input $w(t) = [r_1, r_2, n_3^G]^{\mathrm{T}}$; control $u(t) = [f_{1,2}^{in}, f_{1,3}^{in}, f_{2,2}^{in}, f_{2,3}^{in}]^{\mathrm{T}}$
	$w(1) = [20, 10, 117.5^b]^{\mathrm{T}},\ w(t) = [37, 10, 125^b]^{\mathrm{T}}\ (t \geq 0),\ u(1)^c = [2.5, 7.5, 1.25, 3.75]^{\mathrm{T}}$
Scenario	State vector $x = [(n^H)^{\mathrm{T}}, (n^G)^{\mathrm{T}}, l]^{\mathrm{T}} = [n_1^H, n_2^H, n_0^G, n_1^G, n_2^G, n_3^G, l]^{\mathrm{T}}$
#2	$x(0) = [20, 44, 40^a, 100, 100, 125^b, 30]^{\mathrm{T}}$
	Exogenous input $w(t) = [r_1, r_2, n_3^G]^{\mathrm{T}}$; control $u(t) = [f_{1,2}^{in}, f_{1,3}^{in}, f_{2,2}^{in}, f_{2,3}^{in}]^{\mathrm{T}}$
	$w(1) = [40, 17, 125^b]^{\mathrm{T}}, w(t) = [40, 17, 117.5^b]^{\mathrm{T}}\ (t \geq 0),\ u(1)^c = [2.5, 7.5, 3, 9]^{\mathrm{T}}$

[a] We defined $n_0^G = l + r_1 - f_1^i n$, so here $n_0^G = x_7 + w_1 - u_1 - u_2$.

[b] We defined $n_{N+1}^G = n_{J,N+1}^G - \bar{q}^G = n_{J,N+1}^G - \min[\lambda_N q_{M,N}^G, \hat{w}_{N+1}^G(n_{J,N+1}^G - \hat{n}_{N+1}^G)]$, thus here $n_3^G = n_J - \lambda_2 q_M = 140 - 0.75 \times 30 = 117.5$ means free-flow downstream condition; and $n_3^G = n_J - w(n_J^G - \hat{n}_3^G) = 140 - 0.375 \times (140100) = 125$ means congested downstream condition (i.e., $\hat{n}_3^G = 100 > n_c = 60$).

[c] We assume the initial proportions $P_{1,2} = P_{1,3}$, $P_{2,2} = P_{2,3}$, so $f_{1,2}^{in}/f_{1,3}^{in} = \beta_{1,2}/\beta_{1,3} = 1/3$, $f_{2,2}^{in}/f_{2,3}^{in} = \beta_{2,2}/\beta_{2,3} = 1/3$.

3.5.2 Results and discussions

For this basic modeling segment we use the GLPK solver [1] on a desktop (Intel CPU E5-2680 v3 @3.50 GHz dual processors, RAM 16 G) to compute the optimal solution to the mixed integer linear program (3.42) in an on-line fashion. We simulate the closed-loop performance of the control algorithm for each scenario. Figure 3.5 plots the total cost J and the computational time of the hybrid MPC controller over a period of 20 min (Scenario #1) and 30 min (Scenario #2) under different values of prediction horizon P. Based on (3.40), the total cost is defined as:

$$J = \sum_{t=1}^{T_{tot}} b^{\mathrm{T}} x(t) + \rho 1^{\mathrm{T}} \|u(t) - u(t-1)\|_1, \tag{3.44}$$

where $T_{tot} = 40$ (for Scenario #1) and $T_{tot} = 60$ (for Scenario #2) are the number of time steps over which the controller performance is evaluated. We can see that under both scenarios, the total cost reduces dramatically in the prediction horizon until $P = 4$ (after which the change of cost is negligible). We also observe that the computational time (seconds) increases approximately exponentially in P (i.e., approximately linear in natural log-scale), hence we choose $P = 4$ in our simulation. In particular, under the chosen prediction horizon used by the controller ($P = 4$), the computational times are 1.2 s for Scenario #1 and 2.6 s for Scenario #2, respectively, which are much smaller than the model step of 30 s.

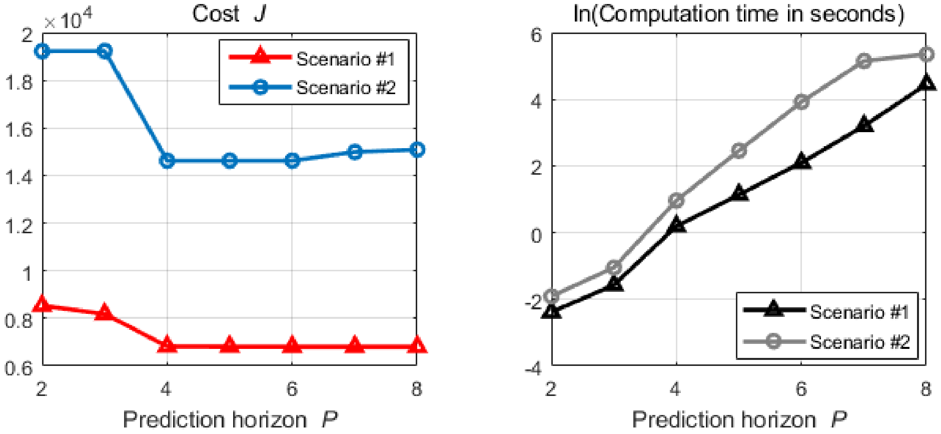


Figure 3.5: Cost and computational time under various prediction horizon.

In Scenario #1, we note that although at $t = 0$ the downstream traffic becomes congested and the upstream demand is nearly doubled, we still have $r_1 + r_2 = 47 < q_M + \bar{q}_3^G = 30 + 20 = 50$, which means there should exits a routing scheme that keep free-flow at HOT lanes while prevent formation of upstream vehicle queues at GP lanes provided no overly stringent constraint. Indeed, we verify that our proposed tolling algorithm can achieve this which successfully found a new equilibrium where vehicle queue is prevented. In Figure 3.6 we plot the evolution of system state, the optimal toll entry flows and the corresponding tolls as well as the active traffic mode on GP lanes. We can see that due to upstream demand increase, overall the tolls $\tau_{1,2}$ and $\tau_{1,3}$ increase significantly to prevent congestion in HOT lanes and the tolls $\tau_{2,2}$ and $\tau_{2,3}$ decrease slightly since the downstream traffic jam forms which reduces the flow that can be supplied by GP lane cell 2. We also observe toll adjustments in the middle of the simulation horizon because our controller has smoothness requirement (3.37) and adjusts its action in a rolling horizon manner, which finds the optimal plan after some time steps. We can see that within 22 time steps (11 min), the controller drives the system from original free-flow state (i.e., mode $k = 11 : \ m_{11} = (D_0, D_1, D_2)$, see Figure 3.4(b)) to a new equilibrium where GP lanes are congested (i.e., mode $k = 6 : \ m_6 = (W_0, W_1, W_2)$). We verify that at the new equilibrium the constraints $P_{2,2} \leq P_{2,2}^{\max}$,$P_{2,3} \leq P_{2,3}^{\max}$ and thus also the constraint (3.29) are binding. We also compare the cost of the fully constrained controller with the one without constraints (3.38) and (3.39), the cost within the 20 min horizon only increases from 8.98×10^3 to 9.04×10^3 and the final throughput per time step are the same (both equal to the maximum possible value 47). Therefore, our controller is optimal for the new exogenous input profile $[r_1, r_2, n_3^G] = [37, 10, 125]$ in the long run. More importantly, it achieves this with both the fairness constraint (3.38) and "no less toll for farther exit" constraint in

(3.39) satisfied, which is crucial for realizing the advantage and feasibility of the OD-based tolling scheme, as discussed in Section 3.3. However, under the new equilibrium derived by the less constrained controller, the constraints $\tau_{1,2} \leq \tau_{1,3}$ and $P_{1,3} \leq 1.2P_{1,2}$ are both violated.

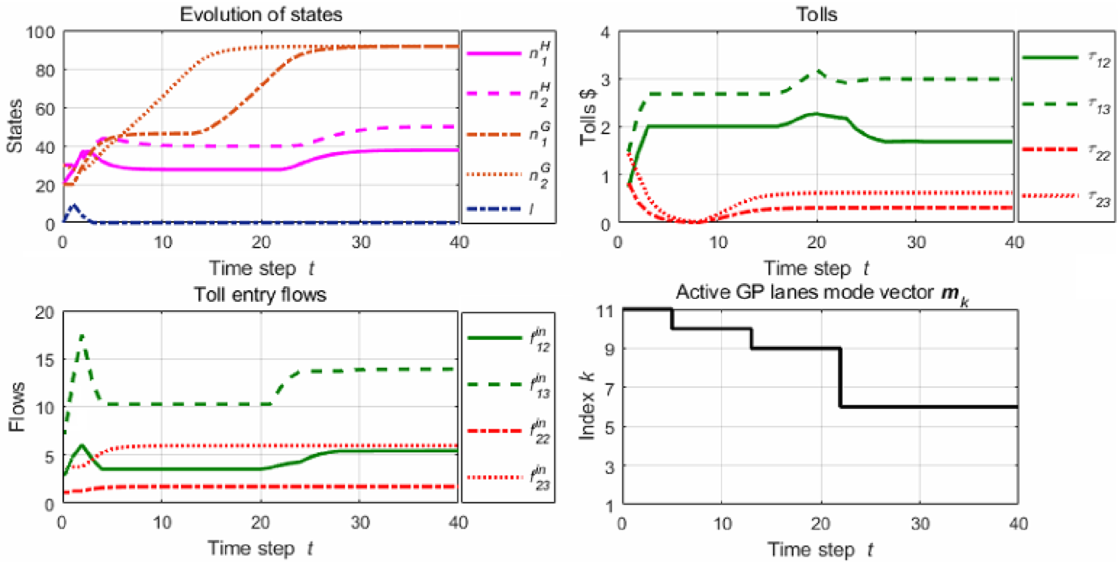


Figure 3.6: Simulation results (Scenario #1).

In Scenario #2, we first verify that under the initial condition the GP lanes are congested with the vehicle queue length increases in a rate of 15 per time step because of the downstream congestion. However, at $t > 0$, the downstream traffic jam disappears and the system has some excess capacity since the total demand $r_1 + r_2 = 57 < 2q_M = 60$. Hence we expect that the system should evolve towards an uncongested condition under dynamic tolling strategy that aims at minimizing the total person delay. Indeed, our controller drives the system to an uncongested equilibrium. Figure 3.7 shows the evolution of system state, the optimal toll entry flows and the corresponding tolls as well as the active traffic mode on GP lanes. We can see that the toll $\tau_{1,2}$ does not change much, but due to the disappearance of the downstream traffic jam, overall the toll $\tau_{1,3}$ that targets the user group with the largest demand (from cell pair 1 to 3) decreases from nearly 3$ to less than 2$ which allows more SOVs enter the first toll entry. In contrast, the tolls $\tau_{2,2}$ and $\tau_{2,3}$ increase in order to maintain free-flow at HOT lanes since the flow at HOT lane cell 1 increases significantly due to notable decrease in $\tau_{1,3}$. We can see that it takes the controller 42 time steps (21 min) to successfully manage the system from original congested mode $m_6 = (W_0, W_1, W_2)$ to the new uncongested mode $m_{11} = (D_0, D_1, D_2)$, which is notably longer than the re-balance time needed for the first scenario since the traffic speed in GP lanes is 24 mile/h under initial congestion, which is much lower than the free-flow speed of 60 mile/h. We verify that at the new equilibrium the only constraint that is binding is $n_2^H \leq n_c^H$, indicating a full utilization of HOT lanes at cell pair 2 that contributes to the improvement in total throughput. Similar with Scenario #1, we notice that the cost within the 30 min horizon increases by 9% (1.46×10^4 versus 1.34×10^4) compared to the cost of the controller that ignores constraints (3.38) and (3.39), under the new flow pattern derived by the less constrained controller, both the constraints $\tau_{2,2} \leq \tau_{2,3}$ and $P_{1,3} \leq 1.2P_{1,2}$ are violated. The final throughput per time step of the two controllers are both equal to the maximum possible value 57. Thus, under the new exogenous input $[r_1, r_2, n_3^G] = [40, 17, 117.5]$, our tolling algorithm achieves the optimal routing plan with both fairness and "no less toll for farther exit" constraints satisfied.

3.6 Conclusions and Future Work

We have developed a hybrid MPC strategy for dynamic tolling of managed lane systems and demonstrated its satisfactory performance on a modeling segment that has four toll entry/exit pairs. The tolling algorithm takes advantage of

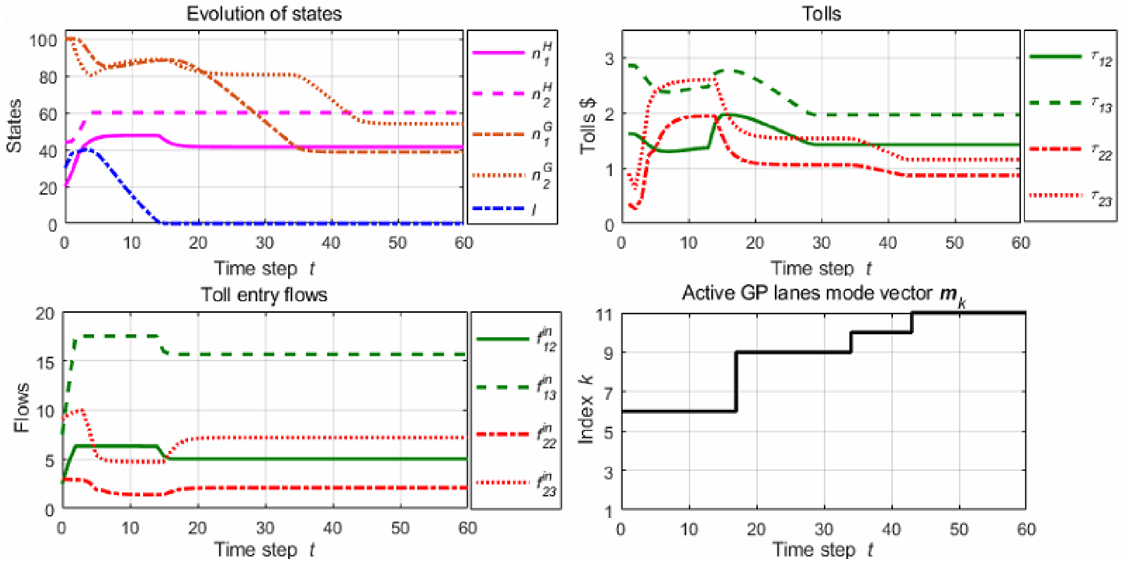


Figure 3.7: Simulation results (Scenario #2).

the traffic demand forecasts and boundary condition measurements and intelligently coordinates the tolls at different HOT lane entries to different downstream exits that collectively minimize the total person delay in the system, including the possible waiting vehicles. Through proper formulation of traffic model and practical constraints, we have shown that the control problem can be cast as a mixed integer linear program which enables both on-line and off-line solution. As shown in the simulation results of the numerical example, the proposed dynamic tolling approach can effectively respond to exogenous input changes by smoothly adjusting the toll entry flows and drive the system to a new optimal state. The proposed control model offers a novel tractable, flexible and nice-structured real-time OD-based dynamic pricing approach for managed lane systems. The general PWA system representation of the traffic flow dynamics in the managed lane systems is also the first of its kind according to our best knowledge.

Further extensions and tests of the proposed hybrid predictive control approach will help bring it closer to real-world implementation. For example, as the size of the optimization problem grows exponentially in the number of cell pairs in the system, the computational cost can be prohibitive when the system is relatively long. Therefore, for large systems a decentralized control model is more suitable (see Remark 3.1), which can be developed based on the proposed model by incorporating additional constraints that impose consistent values at the boundary of two adjacent modeling segments. The control model could be extended to a robust version to deal with error in demand-forecast and off-ramp splitting ratio approximations as well as vehicle-occupancy measurement error (an important issue potentially affecting HOT lane pricing mechanism). In reality, behavior complexities may exist (e.g., relaxation of Assumption 3.4), so the flow distribution and lane-choice models could have variants with time-varying parameters. Thus parameter-learning methods (e.g., for the WTP parameters and the OD ratios) could also be incorporated to achieve better outcomes in practical applications. In addition, efficient coordination with other traffic-control techniques such as ramp metering could be explored, especially when local traffic demand surges.

CHAPTER 4

BAYESIAN OPTIMIZATION FOR SECOND-BEST TOLLING IN TRAFFIC NETWORKS

In this last chapter, we will explore the potential of novel decision-making methodology in solving nontrivial transport network planning problems. As an example, we look at the Second-best Network Pricing Problem (SNPP) in transportation. Recall that in the first chapter we discussed how the intensity of the first-best toll can be reduced using strategic information design. Our focus there was how to utilize information as a “control” measure that helps decentralize an optimal flow using minimal tolls, and the tolls are allowed to be placed anywhere on the network. Now our focus is a commonly adopted “second-best” tolling where only a set of preselected candidate links can be tolled according to real restrictions and needs (such as physical constraints and existing ITS facilities). Unlike the first-best tolling, the second-best network pricing problem is much more complicated, and the problem will become even more challenging if uncertainties such as stochastic demand is considered. To this end, the key in this chapter is designing a solution method that intelligently select and efficiently make use of the information of a few toll scheme “samples” such that we can find a good one as fast as possible.

Specifically, we tailor a Bayesian ranking and selection (R&S) model to solve the SNNP (possibly with demand uncertainty), whose objective is to find an optimal subset of links and toll levels so as to minimize the total travel time on the network. It is in general an NP-hard problem and can have a very large number of candidate solutions. We consider every combination of tollable link(s) and toll levels as an “alternative”, and the problem’s objective function value is regarded as a “reward”, with uncertainties modeled by normal perturbations to the travel demand. We use a linear belief based Knowledge Gradient sampling policy to maximize the expected reward, with Monte Carlo sampling of the hyper-parameters used to reduce the choice set size. Simulation experiments for a benchmark network show the effectiveness of the proposed method and its superior performance to a Sample Average Approximation based Genetic Algorithm.

4.1 Introduction

Network pricing has been widely recognized as an important countermeasure for traffic congestion [129]. One well-known first-best pricing policy involves tolls set at marginal external costs on each link in the network and has been discussed in many studies (e.g., [109, 129]). This policy has been regarded as merely a theoretical construct but impractical for real-world implementation. Under this first-best pricing scheme, total travel time on the network per modeling interval is minimized. Authors of study [16]and [61] solved the problem of finding the first-best pricing scheme that tolls the smallest number of links in a network. They showed that the first-best toll may not be unique. Their optimal “toll set”, however, does not account for restrictions for the location of the tolled links (e.g., restricting tolled links within a certain cordon). Out of practical considerations, most of the recent studies have shifted to solving the second-best network pricing problem (SNPP), e.g., tolling only a subset of links that are tollable. There are generally two branches of research on SNPP: toll level design for a given set of links and optimizing toll rates and link selection simultaneously.

For solving the toll level design problem, derivative-based mathematical programing methods (e.g., [122, 73]) and meta-heuristics such as genetic algorithms (GA) and simulated annealing (SA) (e.g., [131, 111]) have been proposed. Due to path selection assumptions and assignment convergence errors, the derivative-based approaches encounter deficiencies in finding global optimums [111]. It is also known that global optimum is not ensured in meta-heuristics. The authors of study [27] used surrogate optimization method for cordon-based SNPP and achieved close-to-optimal toll solutions with only tens of function evaluations. For joint optimization of toll rates and link selection, three iterative heuristic strategies were proposed in an interesting study [122] based on a "link index" I_a, which represents the welfare gain from implementing a toll on link a alone. This strategy fully accounts for interactions among tolled links but requires calculation of "location indices" for all possible combinations of candidate links. Authors of [110] observed that such an index-based approach has two practical problems: the potential for negative toll predictions and the likelihood of poor initial predictions for parallel links. It was also found that the index-based approach could miss out on toll locations that can yield a high benefit if they are tolled simultaneously [40]. The "location indices" were linked with GA by study [111] to optimize toll locations and used GA to design toll levels given the toll links, due to the "location indices" used, GA often suggested solutions with more toll links that were in fact less optimal. The authors of study [131] used GA for the selection of toll locations and SA for optimal toll level design. Several subsequent studies also applied such heuristics for SNPP (e.g., [140]). However, as is the case for the toll design problem, none of these methods guarantees global optimality [40]. In study [40], the authors instead approximated discrete toll location decision variables with a continuous function and formulated a mixed integer linear program that can be solved for its global optimum. The method only gives a lower bound of the original SNPP, and its computational cost grows rapidly as the accuracy requirement for approximation increases.

All of the aforementioned studies are for deterministic SNPP. Accounting for inherent uncertainty in travel demand, some recent studies (e.g., [50, 77, 117]) started to develop methods that also consider demand uncertainty in SNPP. The discussion in [50] demonstrated better performance of a multiple point inflation/deflation solution method in comparison to that of single point approximation using GA and SA, in terms of computational time versus solution quality. The study was for first-best toll design. Authors of study [77] considered demand uncertainty and environmental externalities for toll-design problems and used sample average approximation (SAA) and sensitivity analysis to solve for the optimal toll levels, given the tolled links. The multi-point approximation method or SAA with a local derivative based method requires a large number (depending on the sample size) of objective function evaluations; and this computational overhead will be costly when the network size is large. In addition, these studies focused on first-best tolling without addressing the more practical second-best toll design, not to mention the consideration of toll location selection. When these toll level design methods are incorporated into heuristics-based toll location optimization problems, the overhead of expensive objective evaluations (simulations) will increase dramatically, making it computationally intractable.

In summary, we feel that there are several important aspects of existing methods for SNPP that need to and can be significantly improved. First, due to the multi-modal nature of the objective functions in SNPP, derivative-based mathematical programming approaches for toll level design can only achieve local optimum. These methods are not suitable for discrete toll location optimization, either. Secondly, heuristic methods, although frequently used for simultaneous toll location and toll level design in SNPP, do not take advantage of the underlying system correlation structure in guiding their search process. Such heuristics generally cannot approach a global optimum within a limited computational budget. In fact, due to the combinatorial nature of toll location plus toll level alternatives, we would expect significant correlations of system performance across candidate solutions that share a common subset of links or that include links on parallel paths for some Origin-Destination (O-D) pairs. Thirdly, in those very limited studies that attempted to also deal with uncertainty in SNPP, the number of scenarios/repetitions used for simulation of candidate solutions was predetermined and fixed. This leads to under- or over-simulation, since it forgoes the opportunity of adjusting the sampling budget dynamically according to the solution quality and the associated uncertainty. Therefore, if we can formulate an SNPP model and design a solution procedure that efficiently leverages upon the underlying correlation structure among different toll levels/locations combinations and their uncertainties, we would be able to improve both the capability and efficiency in approaching or finding the global optimum within limited computational constraints.

With this motivation, in our study we adopt a Bayesian Ranking and Selection (R&S) model and design new solution algorithms to address SNPP for joint optimization of toll locations and toll levels. R&S models [70] have shown superior performance, particularly under a limited sampling budget, in analyzing stochastic outcomes across various alternatives. In the Bayesian R&S formulation, we view each candidate solution to the SNPP as an alternative, and the objective function values are brained by taking “samples” using a Knowledge Gradient policy with Correlated Belief (KGCB) [45]. The Bayesian R&S model fits nicely with SNPP due to its discrete formulation, flexible characterization of correlation structures, and capability to incorporate prior knowledge and the good performance of its sampling policies (e.g., [45, 106]). To adapt the original KGCB sampling strategy to SNPP (which typically has a very large number of alternatives), we further develop the Monte-Carlo-Linear-Belief-KG algorithm (MCLB-KG) based on a non-perfect additive linear belief model to reduce the computational cost for practical SNPP applications.

The rest of this chapter is organized as follows. Section 4.2 introduces the mathematical model of SNPP. Section 4.3 formulates SNPP as a Bayesian R&S problem and describes the construction of the MCLB-KG policy for SNPP. Section 4.4 presents results and discusses computational examples. Section 4.5 concludes.

4.2 Second-Best Network Pricing Problem (SNPP)

Consider a transportation network $G = (N,\ A)$ that consists a set of nodes N and a set of directed links A. There are a set of origin-destination (OD) pairs $D \subseteq N \times N$. There is a traffic demand q_{rs} on OD pair $(r,\ s) \in D$ $(r, s \in N)$. Let K_{rs} be the set of paths starting from node r to node s. We consider a subset of road links, $A' \subseteq A$ with $(|A'| = l)$, be the set of candidate links for pricing (i.e., we are allowed to add tolls to each link $a \in A'$). A' is usually preselected empirically according to congestion level, toll facility installation and operation, existing

ITS facilities, etc. (e.g., [131, 140]). We assume the expected traffic demand is known and inelastic to travel cost.

The SNPP is generally modeled by a bi-level program (e.g., [140, 40]). Let's consider a bi-level program formulation of SNPP with uncertainty. The upper-level problem models the decision maker's objective while the lower-level problem models the network users' travel behavior. The total travel time per time interval on the network is a commonly-adopted measure of traffic efficiency (e.g. [131]). The decision maker's objective is to minimize total travel time per unit time over the network by identifying and tolling a subset of links (selected from predefined candidate links) at appropriate toll levels. Each network user chooses the route with minimum cost to travel from her origin to her destination. Assuming homogeneous unit "value of time" (VOT) among users, the formulation of SNPP is given as:

$$\text{(Upper} - \text{level)} \qquad \max_{d} \mathbb{E}(T_d) = \int_{\omega\in\Omega} \left[T_0(\omega) - \sum_{a\in A} z_a^*(\omega) t_a(z_a^*(\omega)) \right] p(\omega) d\omega \tag{4.1}$$

$$\text{s.t. } d = [d_1, ..., d_l]^{\mathrm{T}},\ d_i \in \{0, 1, ..., m\},\ \forall i \in A'$$

$$\text{(Lower} - \text{level)} \qquad \min_{z} Z(d) = \sum_{a\in A\setminus A'} \int_0^{z_a} t_a(v)dv + \sum_{a\in A'} \int_0^{z_a} [t_a(v) + u_a]dv \tag{4.2}$$

$$\text{s.t. } \sum_{k} f_{rs}^k = q_{rs},\ f_{rs}^k \geq 0,\ \forall (r,s) \in D,\ \forall k \in K_{rs}$$

$$z_a = \sum_{(r,s)\in D} \sum_{k\in K_{rs}} f_{rs}^k \delta_{rs}^{ak}.$$

In the upper level problem, we maximize $\mathbb{E}(T_d)$, the expected difference between the total travel time of the no-toll scenario, T_0, and the total travel time of the tolled scenario, $\sum a \in A z_a^* t_a(z_a^*)$. z_a^* is the traffic volume per unit time on link a under the optimal solution of the lower level problem; t_a is the corresponding travel time on link a, which is a function of z_a^*. In the objective function, uncertainty is generally considered by defining a random variable ω that lies in a known space Ω. For generality, we assume ω is continuous (which can also be used to model countable scenario setting, e.g., [50]). Model inputs and parameters such as traffic demand can take on random values. The probability density function of ω is $p(\omega)$ and assumed to be known. Traffic demand plays a fundamental role determining network performance. Hence we focus on random demand for uncertainty consideration in SNPP. We use ω to represent the stochasticity in OD demand $\{q_{rs}\}$, and Ω is the set of possible outcomes of ω related to the demand. $d = (d_1, ..., d_l)^{\mathrm{T}}$ is an integer-valued decision vector, i.e., $d_i \in \{0, 1, ..., m\}$, $i = 1, ..., l$, indicating the possible toll levels to be applied to each candidate link. For example, if $m = 3$, then $d_i = 0, 1, 2, 3$ represent no toll, low, medium and high toll levels, respectively. Link travel time $t_a(z_a)$ is a non-decreasing convex function of traffic volume z_a, and the popular BPR formula [97] is used here: $t_a = t_a^0[1 + \alpha(z_a/c_a)^\beta]$, where $\alpha > 0$, $\beta > 1$ are empirical parameters, t_a^0, c_a are free-flow travel time and capacity of link a, respectively. $u_a = e \cdot d_i/\mathrm{VOT}$ (i corresponds to a) is the equivalent time cost of the toll on a candidate link $a \in A'$, e is the unit toll level.

Given candidate link set A', number of toll levels m and incremental unit e across toll levels, the key inputs to this upper-level maximization problem of SNPP is the traffic flow distribution on the all the directed links of the network, $z* = \{z_a*\}$, resulting from the solution of the lower-level problem. The link flow distribution $\{z_a*\}$ is the result of the traffic flow assignments represented by the flows on the set of paths $k \in K_{rs}$ for each OD pair $(r, s) \in D$, $\{f_{rs}^k\}$, which are part of the variables to the lower level problem. The binary variable δ_{rs}^{ak} takes value one if path $k \in K_{rs}$ contains link a, and zero otherwise. The lower-level problem is to find a user equilibrium (UE) flow pattern with potential equivalent time cost u_a considering link travel time and toll; under UE no user has incentive to change route. The UE problem can be formulated as the form of (4.2) and has well-established solution methods like Frank-Wolfe algorithm [109], but the computational cost grows significantly with the network size due to the shortest path subroutine involved. Since the lower- and upper- level problems in SNPP are hard to be integrated as one objective due to intrinsic difficulty of the problem (e.g., [131]), the SNPP can be regarded as a "black-box" discrete optimization problem. Furthermore, if considering more complicated case such as stochastic UE [109], the lower-level problem may not have equivalent mathematical programing formulation such as the one in (4.2), in which case it can only be solved by numerical or simulation approaches, which is in general costly. This nature of SNPP is at the heart of its Bayesian R&S formulation. In this study, we focus on the simple case where lower level UE problem can be solved by solving (4.2) as a illustration.

4.3 SNPP as a Bayesian R&S Problem with Linear Beliefs

4.3.1 Bayesian R&S formulation of SNPP

In a Bayesian R&S framework, we have M alternative decisions $X = \{x_1, x_2, ..., x_M\}$ whose rewards (e.g., the values of objective functions for different pricing schemes in SNPP) are random with unknown mean $\theta = (\theta_1, ..., \theta_M)^{\mathrm{T}}$ and unknown variance $\lambda = (\lambda_1, ..., \lambda_M)^{\mathrm{T}}$. Our goal is to identify the alternative with the maximum expected reward through limited sample measurements. We have a prior belief about θ with mean $\mu^0 = \{\mu_1^0, ..., \mu_M^0\}$ and covariance Σ^0 (an $M \times M$ positive semi-definite covariance matrix). For SNPP, we have network performance under different pricing schemes as $\theta = (\mathbb{E}(T_{d_1}), ..., \mathbb{E}(T_{d_M})]^{\mathrm{T}} \sim \mathcal{N}(\mu^0,\ \Sigma^0)$. Assume that we can evaluate N sample decisions, $x^0,\ x^1, ..., x^{N-1}$. At stage $n = 0, ..., N-1$, we make a measurement or evaluation of decision x^n, with measurement noise, $\epsilon^n \sim \mathcal{N}(0,\ \lambda_{x^n})$, independent across samples conditional on x^n. This yields sample observation (i.e., objective function evaluation in SNPP) $y^{n+1} = \theta_{x^n} + \epsilon^n$. Let F^n be the sigma-algebra generated by $\{x^0, ..., x^{n-1}\}$ and $\{y1, ..., y^n\}$. It is a well-known result that the conditional posterior distribution of θ is also multivariate normal. Let $\mu^n = \mathbb{E}(\theta|F^n)$ and $\Sigma^n = Cov(\theta|F^n)$ be stage-n conditional expectation and covariance of θ, respectively, then μ^n and Σ^n can be calculated recursively using standard results based on Bayes' Rule (e.g., [51]):

$$\mu^{n+1} = \mu^n + \frac{y^{n+1} - \mu^n_{x^n}}{\Sigma^n_{x^n x^n} + \lambda_{x^n}} \Sigma^n e_{x^n},\ \Sigma^{n+1} = \Sigma^n - \frac{\Sigma^n e_{x^n} e^{\mathrm{T}}_{x^n} \Sigma^n}{\Sigma^n_{x^n x^n} + \lambda_{x^n}}, \tag{4.3}$$

where e_{x^n} is a column vector of zeros with a one at the entry corresponds to x^n. $\Sigma^n_{x^n x^n}$ is the diagonal entry of matrix Σ^n corresponds to x^n. λ_{x^n} is the performance variance corresponding to decision alternative x^n.

After N measurements through a sampling policy $\pi = \{x^1_\pi, ..., x^N_\pi\}$ from the policy space Π, we choose the alternative that yields the largest posterior mean of objective function value (rewards) as the optimal solution: $\sup_{\pi\in\Pi} \mathbb{E}^\pi[\max_x(\mu^N_x)]$, where $\mathbb{E}^\pi$ denotes the conditional expectation under π. In Bayesian R&S formulation of SNPP, as we evaluate pricing alternatives $d_{x^1}, ..., d_{x^N}$, we obtain measurements of the random "rewards" which represents total network travel time reductions $T_{d^1}, ..., T_{d^N}$ in comparison to the no-toll baseline scenario.

4.3.2 KGCB sampling policy

The Knowledge Gradient policy with Correlated Belief (KGCB policy) is originally introduced in [45]. It samples alternative x that maximizes the incremental value (knowledge gradient) of the objective function:

$$x^{KG,n}(S^n) = \arg\max_x v^{KG,n}(x) = \arg\max_x \left(\mathbb{E}^n[\max_i \mu_i^{n+1} | S^n,\ x^n = x] - \max_i \mu_i^n\right), \quad (4.4)$$

where $S^n = (\mu^n, \Sigma^n)$ is the state of our posterior beliefs at measurement n. The KG-factor $v^{KG,n}(x)$ represents the incremental value (i.e., the expected improvement in posterior optimal value) obtained from measuring x at stage n. It is shown that the KGCB policy is almost-surely optimal for $N = 1$ or $N \to \infty$, and has sub-optimality bounds when N is finite [45].

By calculating the conditional predictive expectation of $\max_x \mu_x^{n+1}$, we can forecast the performance of all alternatives without taking actual samples of them. Therefore, one key step in KGCB policy is to compute conditional predictive distribution of $^{n+1}$ given information at stage n. This conditional distribution is multivariate normal, with mean $\mathbb{E}^n[\mu^{n+1}] = \mu^n$ and covariance $\tilde{\sigma}(\Sigma^n,\ x^n)\tilde{\sigma}(\Sigma^n,\ x^n)^{\mathrm{T}}$, where $\tilde{\sigma}(\Sigma,\ x) = \Sigma^n e_x / \sqrt{\lambda_x + \Sigma_{xx}^n}$, details of this calculation can be found in [45]. Thus the stage-n conditional distribution of μ^{n+1} is the same as $\mu^n + \tilde{\sigma}(\Sigma^n,\ x^n)Z$, where Z is scalar standard normal random variable. This allows us to compute $v^{KG,n}(x)$ in (4.4) as

$$v^{KG,n}(x) = E^n[\max_i \mu_i^n + \tilde{\sigma}(\Sigma^n,\ x^n)Z | S^n,\ x^n = x] - \max_i \mu_i^n = h(\mu^n,\ \tilde{\sigma}(\Sigma^n,\ x^n)), \quad (4.5)$$

where function $h : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ is defined as $h(p,\ q) = \mathbb{E}[\max_i p_i + q_i Z] - \max_i p_i$, here p and q are deterministic M-dimensional vectors. A method of computing $h(p, q)$ is presented in [45], where the entries of p and q are firstly sorted and then only the distinct ones retained to define a vector c with $c_i = (p_i - p_{i+1})/(q_{i+1} - q_i)$.

These quantities are then used to calculate $h(p, q) = \sum_{i=1,...,\dim(p)-1}(q_{i+1} - q_i)f(-|c_i|)$, where $f(z) = \Phi(z) + z\phi(z)$, Φ and ϕ are standard normal PDF and CDF, respectively. We call this method "Subroutine-*h*", by which we can compute $h()$ for any prior belief μ and $\tilde{\sigma}(\Sigma, x)$. Then we are able to compute $v^{KG,n}(x)$ for each alternative x, and the largest $v^{KG,n}(x)$ gives the measurement of decision $x^{KG,n}$.

In the standard setting of KGCB, the dimensions of q, p and μ^n are the number of alternatives, M. Therefore, the Subroutine-*h* is executed M times for obtaining $x^{KG,n}$. Since the sorting step dominates the computational cost of Subroutine-*h*, so the overall complexity of the standard KGCB algorithm is $O(M^2 \log M)$. Thus for large number of alternatives, say $M > 10^5$ (which is usually the case for SNPP due to large number of link/toll combinations), the computational demand of the standard KGCB is prohibitive. This leads us to the modification of KGCB as follows.

4.3.3 Linear belief model for SNPP

In SNPP, the compounding effect of tolls on multiple links at their respective toll levels is not simply additive, i.e., summation of individual "link indices" of links a and b, $I_a + I_b$, will not simply be equal "location index", I_{ab}, the effect of simultaneous tolls on links a and b [122]. In fact, the interaction effects among tolled links, although hard to quantify, can be remarkable, especially among links on parallel paths for the same OD pair [110]. Therefore, we propose a non-perfect additive linear belief model to consider such joint effect from tolling multiple links. This approach is inspired by the linear belief model used by [94] in their study of sequential experimenting for drug discovery.

Model structure and priors on model coefficients

For our Bayesian R&S SNPP, we assume the marginal effect (on the final objective value) from one unit increase in toll rate on a link varies significantly across different toll levels. Thus we have $m \times l$ attributes (l candidate links, each with m toll levels). This leads to a new binary column decision vector d^e of size ml, expanded from the original l-dimension decision vector d. By assigning m entries for each candidate link j in set A' and placing these m-entries across the links in the order of $j = 1, 2, ..., l$, we have:

$$d_i^e = \begin{cases} 1, & [i - m(j-1)]^{\text{th}} \text{ toll level onlink } j \\ 0, & \text{otherwise} \end{cases} \quad \forall i = m(j-1)+1, ..., mj, \; j = 1, ..., l. \tag{4.6}$$

For example, consider a toy example with only two candidate links (dimensions) and two toll levels (attributes), i.e., $l = |A'| = 2$, $m = 2$, $j \in \{1, 2\}$. Using the notation above, the no-toll alternative can be represented by $d^e_{(00)} = (0, 0, 0, 0)^{\text{T}}$, where the first two zeros correspond to first link and the last two for the second link. The alternative of applying toll level 2 on link 1 and toll level 1on link 2 is represented by $d^e_{(21)} = (0, 1, 1, 0)^{\text{T}}$.

We now assume a non-perfect linear additive model for the effect of a SNPP pricing scheme d_x:

$$\theta_x = \eta_0 + \sum_{i=1}^{ml} \eta_i d^e_{x,i} + \zeta_x, \tag{4.7}$$

where η_0 is the value for the no toll case (i.e., all entries in d^e_x are zero); coefficient η_i represents the marginal effect per unit change in attribute d^e_i (here it is the $[im(j-1)]^{\text{th}}$ toll level on link j); ζ_x is the deviation term from perfect additive structure, which is alternative specific as labeled by subscript x.

This non-perfect linear additive model is similar to that in [94]. It is generalized from the perfect linear-additive model, Free-Wilson model [46], by adding the deviation term ζ_x to the performance of alternative x. Since only one toll level is to be implemented for each candidate link (as we focus on static network optimization), i.e., $\sum_{m(j-1)<i\leq mj} d_i^e = 1$, so the requirement of the Free-Wilson model that "at most one attribute is associated with each dimension" is automatically satisfied. Based on this linear belief model, if we sample d_x^e (corresponding to d_x), the sample value would be

$$T_{d_x} = y_x = \eta_0 + \sum_{i=1}^{ml} \eta_i d_{x,i}^e + \zeta_x + \epsilon_x, \tag{4.8}$$

where $\epsilon_x \sim \mathcal{N}(0,\ \lambda_x)$ is an independent measurement noise for d_x. $\lambda_x = 0$ models the deterministic SNPP, and $\lambda_x > 0$ addresses SNPP with uncertainty (due to stochastic demand in our case).

Suppose we have independent normal priors on η_0 and $\eta_i,\ i = 1, ..., ml$: $\eta_0 \sim \mathcal{N}(\mu_{\eta_0},\ \sigma_{\eta_0}^2)$, $\eta_i \sim \mathcal{N}(\mu_{\eta_i},\ \sigma_{\eta_i}^2),\ i = 1, ..., ml$. To begin with, we use independent normal distributions with mean 0 and variance σ_ζ^2 as priors for $\zeta_1, ..., \zeta_M$. Note that $\zeta_1, ..., \zeta_M$ are independent from other model coefficients. Then the prior belief about the expected value of decision d_x is

$$\mu_x^0 = \mu_{\eta_0} + \sum_{i=1}^{ml} \mu_{\eta_i} d_{x,i}^e, \tag{4.9}$$

and the prior belief of the covariance between the performance of d_x and d_x' is [94]

$$\begin{aligned}\Sigma^0(x,\ x') &= \operatorname{cov}(\eta_0 + \sum_{i=1}^{ml} \eta_i d_{x,i}^e + \zeta_x,\ \eta_0 + \sum_{i=1}^{ml} \eta_i d_{x',i}^e + \zeta_{x'} \\ &= \sigma_{\eta_0}^2 + \sum_{i=1}^{ml} d_{x,i}^e d_{x',i}^e \sigma_{\eta_i}^2 + \sigma_\zeta^2 \mathbf{1}_{\{x=x'\}},\end{aligned} \tag{4.10}$$

where $\mathbf{1}_{\{\}}$ is the indicator function.

Posterior distributions on model coefficients

Maintaining and updating posteriors on linear belief model coefficients (marginal effects of different attributes) is a key step in solving the Bayesian R&S SNPP. Let the column coefficient vector η denote $(\eta_0, \eta_1, ..., \eta_{ml})^{\mathrm{T}}$ and $D_{M\times(ml+1)}$ be a matrix comprised of rows each representing the attribute values of an alternative plus a "1" in the first entry corresponding to the baseline (no-toll scheme) constant η_0. Thus a row in D is a "1" followed by the attribute values of d^e. We also use a column vector ζ to denote all the deviation terms $\{\zeta_x\}$. With these notations, the true value vector is $\theta = D\eta + \zeta$ by (4.7). Even though the number of ζ_x is M, which is generally very large in SNPP, we only need to maintain a mean vector and covariance matrix of ζ_x's for alternatives that have already been measured. If we have not measured a alternative x by stage n, then the posterior of ζ_x will stay the same as its prior. ζ_x remains independent of η_i's, and of all the other deviation terms.

Toward this end, we define column vector η^n that contains η and ζ_x terms for an alternative x in $\{x^0, ..., x^{n-1}\}$. Let a^n and C^n be the mean vector and covariance matrix of our stage-n posterior of η^n. Note that before the first measurement, the initial values are $\eta^0 = \eta$, $a^0 \in \mathbb{R}^{(ml+1)\times 1} = \{\mu_{\eta_i},\ i = 0, ..., ml\}$, and diagonal matrix $C^0 \in \mathbb{R}^{(ml+1)\times(ml+1)}$ with diagonal entries $\{\sigma^2_{\eta_i},\ i = 0, ..., ml\}$. Due to the property of multivariate normal distribution, there is a recursive expression for a^n and C^n [94]

$$a^{n+1} = \tilde{a}^n + \frac{y^n - (\tilde{a}^n)^{\mathrm{T}}\tilde{d}_{x^n}}{\lambda_{x^n} + (\tilde{d}_{x^n})^{\mathrm{T}}\tilde{C}^n\tilde{d}_{x^n}}\tilde{C}^n\tilde{d}_{x^n}, \tag{4.11}$$

$$\tilde{C}^{n+1} = \tilde{C}^n - \frac{\tilde{C}^n\tilde{d}_{x^n}(\tilde{d}_{x^n})^{\mathrm{T}}\tilde{C}^n}{\lambda_{x^n} + (\tilde{d}_{x^n})^{\mathrm{T}}\tilde{C}^n\tilde{d}_{x^n}}, \tag{4.12}$$

where $\tilde{a}^{n-1}$ and $\tilde{C}^{n-1}$ are defined as below: if x^n has been previously measured by stage n, $\tilde{a}^{n-1} = a^{n-1}$, $\tilde{C}^{n-1} = C^{n-1}$; otherwise, let $\tilde{a}^{n-1}$ be the column vector formed by appending a 0 to the bottom of a^{n-1}, and $\tilde{C}^{n-1}$ be the matrix formed by adding a row and a column beneath the bottom row and to the right of the rightmost column of C^{n-1}, where all the entries of the new row and the new column are zero but the diagonal entry is σ_{ζ}^2 (i.e., the rightmost bottom entry is σ_{ζ}^2). Then the posterior of η^n at stage $n - 1$ is $\mathcal{N}(\tilde{a}^{n-1}, \tilde{c}^{n-1})$. $\tilde{d}_{x^n}$ is a column vector consisting of 1′s at indices of η^{n+1} for which alternative x^n contains the corresponding baseline term, toll level attributes and deviation term, and zero elsewhere. (4.11)-(4.12) can be viewed as a linear square recursive model (e.g., [102]) modified by incorporating the deviation terms from our non-perfect linear additive model for SNPP. These updating equations allow us to track and update our beliefs on η^n in a computationally efficient way. The prior beliefs of the model coefficients, parameterized by μ_{η_0}, $\sigma_{\eta_0}^2$, μ_{η_i}, $\sigma_{\eta_i}^2$, σ_{ζ}^2 can be estimated from initial sampling or prior information.

Based on the updated beliefs of the hyper-parameters, we can construct the posteriors of the alternatives' values. Noting that any multivariate normal belief on η^n induces a multivariate normal belief on θ^n [94], we have $\theta^n \sim \mathcal{N}(\mu^n, \Sigma^n)$ from $\eta^n \sim \mathcal{N}(a^n, C^n)$. μ^n and Σ^n are calculated from a^n and C^n using the same method as (4.9) and (4.10). However, to use KGCB algorithm, we only need to retrieve partial information without computing the whole Σ^n matrix (which is prohibitive in SNPP).

4.3.4 Updating the unknown variances

For the SNPP under demand uncertainty, the variance for each alternative x, λ_x (i.e., the variance of measurement noise ϵ_x in (4.8) is usually unknown. With very limited sampling budget, this variance affects the belief update of the hyper-parameters in (4.11)-(4.12) and the characterization of conditional distribution of μ^{n+1}. Therefore, an estimation updating procedure of λ_x is needed to improve the performance of sampling policy. We use an approach inspired by the Bayesian normal model with known mean and unknown variance [51] to estimate λ_x. We start with a prior belief λ_x^0 that is constant or varying across alternatives, it can be simply the best guess based on the information available. As the learning progresses in implementing the solution algorithm, we can collect more samples for a certain decision vector d_x and update our estimate of that λ_x. In iteration n where alternative x is sampled, we use

$$\begin{aligned}\lambda_x^n = \frac{\lambda_x^{n-1}(w + ns_x^n - 3) + (y_x^n - \mu_x^n)^2}{w + ns_x^n - 2} &= \frac{w\lambda_x^0 + \sum_{i=1}^{ns_x^n}(y_x^{n_x(i)} - \mu_x^{n_x(i)})^2}{w + ns_x^n - 2} \\ &\approx \frac{w\lambda_x^0 + \sum_{i=1}^{ns_x^n}(y_x^{n_x(i)} - \theta_x)^2}{w + ns_x^n - 2},\end{aligned} \tag{4.13}$$

where ns_x^n is the x^{th} entry in the M-dimensional vector ns^n used to record how many times each alternative has been sampled up to stage n; iteration $n_x(i)$ is the iteration when alternative x is sampled for the i^{th} time; $\mu_x^{n(i)}$ is the posterior mean of x at iteration $n_x(i)$; and weight $w \geq 0$. The idea behind (4.13) is to estimate λ_x as a weighted average of the prior belief and the information observed by the samples. To marginalize the impact of inaccuracy from posterior means, we require $n \geq 3$ before (4.13) is applied. As the number of samples increases, the variance estimates will gradually converge to their true values.

4.3.5 KGCB Algorithm for SNPP with linear beliefs and MC sampling for the hyper-parameters

Now we can compute the KG factors from a belief on η^n parameterized by a^n and C^n. By (4.5), we can obtain $v^{KG,n}(x)$ using Subroutine-h when μ^n and $\tilde{\sigma}(\Sigma^n, \ x)$ are available. Independent of x, $\mu^n = D^n a$, where D^n is a $M \times \dim(^n)$ indexing matrix of 0's and 1's, each row of which corresponds to an alternative and has 1's for the baseline term, toll attribute terms and the deviation terms from a^n that are contained in the alternative. To compute $\tilde{\sigma}(\Sigma^n, \ x)$, we set $x^n = x$ and get the corresponding η^{n+1} and $\tilde{C}^n$. Let $\tilde{D}^n$ be a $M \times \dim(\eta^{n+1})$ matrix that is similar to D^n, except that it maps alternatives to component of η^{n+1} instead of η^n. Note that the beliefs on those ζ_x terms that are not included in η^{n+1} will not change as a result of measuring x^n. In addition, $\tilde{\sigma}(\Sigma^n, \ x^n)$ is not affected by such deviation terms. So we can ignore the left-out deviation terms, by the derivation in [45], $\tilde{\sigma}(\Sigma^n, \ x) = \Sigma^n_{\cdot,x} / \sqrt{\lambda_x + \Sigma^n_{xx}}$, where $\Sigma_{\cdot,x}$ denotes the x^{th} column of Σ^n, which equals

$$\Sigma^n_{\cdot,x} = \tilde{D}^n \tilde{C}^n (\tilde{D}^n_{\cdot,x})^{\mathrm{T}}. \tag{4.14}$$

Then we can compute the KG factor $v^{KG,n}(x)$ for all decisions x. The standard KGCB algorithm based on our non-perfect additive linear belief model is summarized below, we refer to it as the LB-KG Algorithm (where $\lambda^n = \{\lambda^n_x\}$). The complexity of the LB-KG Algorithm is $O(M^2 \log M)$.

However, when M is large, computing KG factors for all decisions as required in standard KGCB algorithm is very expensive. Inspired by [106], we propose a Monte Carlo (MC) sampling step to substantially reduce the size of the choice set. But instead of sampling θ from $\mathcal{N}(\mu, \ \Sigma)$ as used in [106], we directly sample from hyper-parameter space and generate realizations of θ accord-

Algorithm 4.1: LB-KG Algorithm (stage n)

```
Require: $M$, $\lambda^n$, $D^n$, $a^n$ and $C^n$
$\mu^n \leftarrow D^n a^n$, $v^* = 0$
for $x = 1$ to $M$ do
    Compute $\tilde{C}^n$ from $x$ and $C^n$
    $\Sigma^n_{\cdot,x} \leftarrow \tilde{D}^n \tilde{C}^n (\tilde{D}^n_{\cdot,x})^{\mathrm{T}}$
    $p \leftarrow \mu^n,\ q \leftarrow \Sigma^n_{\cdot,x} / \sqrt{\lambda_x + \Sigma^n_{xx}}$
    $v \leftarrow h(p,\ q)$ using Subroutine-$h$ (described in Section 4.3.2)
    if $x = 1$ or $v > v^*$ then
        $x^* \leftarrow x$, $v^* \leftarrow v$
    end if
end for
return: $x^*$
```

ing to the linear belief model, which in the first place permits significant savings. Suppose we generate K sample realizations of $\bar{\theta}^n$ based on the non-perfect linear additive models and the posterior beliefs of the model coefficients at stage n. Let $\bar{\eta}^n(\omega_k)$ be the k^{th} sample realization of model coefficients from the posterior distribution $\mathcal{N}(a^n,\ C^n)$. The M-dimensional column vector $\bar{\zeta}^n(\omega_k)$ has zero entries for sampled alternatives, and each entry of $\bar{\zeta}^n(\omega_k)$ corresponds to unmeasured alternatives is separately drawn from the prior distribution $\mathcal{N}(0,\ \sigma^2_\zeta)$. Then the mean $\bar{\theta}^n(\omega_k)$ of the k^{th} sample realization will be an M-dimensional column vector for each k

$$\bar{\theta}^n(\omega_k) = D^n \bar{\eta}^n(\omega_k) + \bar{\zeta}^n(\omega_k). \tag{4.15}$$

Lett_k $= \arg\max_t \bar{\theta}^n_t(\omega_k)$ be the toll alternative that appears to be the best from sample k and let K_0 be the number of such distinct alternatives from all K samples. The number of alternatives in SNPP is much larger than that was encountered in [106], so as a remedy, in iteration n, we propose to randomly sample K_1 distinct alternatives $s_1, ..., s_{K_1}$ from the complete candidate solution space and use the final choice set $S = \{t_1, ..., t_{K_0}\} \cup \{s_1, .., s_{K_1}\}$. Then the KG-factors $v^{KG,n}$ can be computed only over set S. We call this the Monte Carlo linear belief KG policy (MCLB-KG), which is adopted for our challenging SNPP. Thus the complexity of MCLB-KG algorithm in v^{KG} calculation is $O(|S|^2 \log |S|)$, much less than $O(M^2 \log M)$ in the standard KGCB policy (Algorithm 4.1. Also we only need to sample a vector with dimension equal to $\dim(\eta^n)$ (note that $\dim(\eta^n) \leq n + lm + 1$) from multivariate normal distribution $\mathcal{N}(a^n,\ C^n)$ at iteration n, the complexity of the MC sampling step is $O(\dim(\eta^n)^3 K)$ when the Cholesky factorization of C^n is used (which is very efficient in modern computing package), so this implementation is $<< O(|S|^2 \log |S|)$. Note that recognizing those unmeasured alternatives by stage n can be done efficiently by keeping a list of sampled alternatives rather than looping over tags for M alternatives. Hence the other overhead of the MCLB-KG algorithm mainly comes from the multiplications of high-dimension matrices in (4.14) and (4.15), which have only linear dependency on M. Therefore, the computational cost can be significantly reduced compared to the LB-KG Algorithm. We summarize this MCLB-KG Algorithm below.

Algorithm 4.2: MCLB-KG Algorithm (stage n)

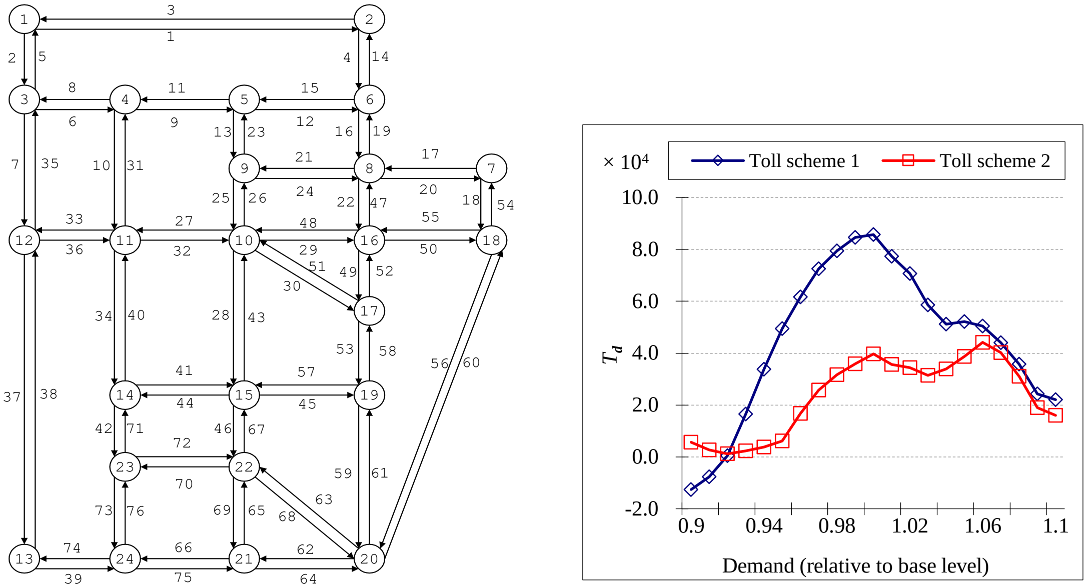
```
Require: $M, K, K_1, ns^n, \sigma_\zeta^2, \lambda^n, D^n, a^n$ and $C^n$
$S \leftarrow \emptyset$
for $k = 1$ to $K$ do
    Draw a MC sample $\eta_n(\omega_k) \sim \mathcal{N}(a^n,\ C^n)$
    $\bar{\zeta}^n(\omega_k) \leftarrow \mathbf{0}$
    temp $\leftarrow$ vector that contains $M$ MC samples of $\mathcal{N}(0,\ \sigma_\zeta^2)$
    for $x = 1$ to $M$ do
        if $ns^n(x)$ =0 then
            $\bar{\zeta}_x^n(\omega_k) \leftarrow temp(x)$
        end if
    end for
    Calculate $\bar{\theta}^n(\omega_k)$ by (4.15)
    $t \leftarrow \arg\max_x \bar{\theta}_x^n(\omega_k)$
    if $t \notin S$ then
        $S \leftarrow S \cup \{t\}$
    end if
end for
for $k = 1$ to $K_1$ do
    Choose a random integer $s$ from $\{1, 2, ..., M\}$
    if $s \notin S$ then
        $S \leftarrow S \cup \{s\}$
    end if
end for
Compute $x^*$ from all $x \in S$ by Algorithm 4.1 with $|S|$ as input instead of $M$
```

4.4 Numerical Experiment

4.4.1 Input and simulation data

We apply the method to the benchmark Sioux Falls network, which is used in recent SNPP studies (e.g., [40]). It has 24 nodes and 76 links and 576 OD pairs (see Figure 4.1(a)) with detailed network date given in [10]. Due to budget constraints, 10 candidate links $A' = \{16, 19, 29, 39, 48, 49, 52, 66, 74, 75\}$ based on initial congestion levels are of interest, 3 toll levels are proposed with unit toll level $e = \$2$. The homogeneous VOT = \$1/min. The total travel time under base demand is $T_0 = 8 \times 10^6$ min per unit time. In the implementation of MCLB-KG policy, we set the number of random samples $K = 100$ and $K_1 = 200$. We use non-informative priors for most of the parameters in the belief model: $\sigma^2_\zeta = 10^5$, $\sigma^2_{\eta_i} = 4\times10^5$, $\mu_{\eta_i} = 400,\ 800,\ 1200$ for toll levels 1, 2 and 3, respectively. Although the prior means of the toll attributes' marginal effects are positive, large uncertainties are attached to these coefficients as well as to the deviation terms. We have almost complete information about the baseline no-toll alternative, so we set $\mu_{\eta_0} = 0$ and $\sigma^2_{\eta_0} = 10$ for deterministic tests and $\sigma^2_{\eta_0} = 10^4$ for stochastic case. We use a non-informative prior to demonstrate the effective learning capability of the MCLB-KG policy for SNPP.

We examine the performance of Bayesian R&S SNPP model solved via MCLB-KG algorithms in comparison to the GA (which is usually used for SNPP) for deterministic setting ($\lambda_x = 0$) and SAA-GA for stochastic ($\lambda_x > 0$) setting. We use the standard GA [35] with population size $|A'|$ and elite size 1 (optimized by grid search). SAA is used for GA to evaluate individual solution (e.g., [50]) with sample size 5 for stochastic setting (performs best among 2~6).



(a) Network layout

(b) Travel time reduction under two toll alternatives

Figure 4.1: Sioux Falls test network and non-normality of alternative values.

The simulation budget N is 100 and 300 for the deterministic and stochastic tests, respectively. Because in stochastic case, evaluation of one solution contains two simulations under the same demand realization, one for the toll alternative and the other for the non-toll baseline case, and the SAA sample size is 5 for GA, so this means 100 and 150 iterations in R&S and $\lceil 100/|A'|\rceil$ and $\lceil 30/|A'|\rceil$ generations in GA (or SAA-GA) for the deterministic and stochastic cases, respectively. In stochastic test, each OD demand q_{rs} in (4.2) is subject to a common $p\% \sim \mathcal{N}(0, v^2)$ perturbation, q_{rs} is set to 0 if it drops below 0. Two cases $v = 0.01$ and 0.05 are tested, with $\lambda_x^0 = 10^5$ and 4×10^5, respectively. We run 10 independent sample paths for each algorithm in both deterministic and stochastic tests. The main performance measure is the Relative Opportunity Cost (RelOC), defined as the relative difference between the objective value of the true optimal solution (the best solution possibly known) and the objective value of the best alternative proposed by the algorithm.

Note that normally distributed perturbation in traffic demand does not necessarily results in normally distributed objective value T_d, as shown in Figure 4.1(b). We can see the travel time reductions under two toll alternatives are not affine in demand with markedly different patterns. This is due to the nonlinear function $t_a(z_a)$ and complicated system response of underlying UE flow assignment. We use this deviation from normality to show the robustness of the normality based Bayesian R&S algorithm for practical problems such as SNPP. This also indicates that the objective value evaluated at base demand may not be the true value of an toll alternative. Thus the mean of 1000 Monte Carlo samples is used as this "true" value.

4.4.2 Results and discussion

Figure 4.2 (a), (b) and (c) compare the RelOC between Bayesian R&S SNPP solved by MCLB-KG and solved by GA or SAA-GA (point estimate of each RelOC with its ~95% confidence interval (CI) plotted as "error bars"). In all three cases, the MCLB-KG algorithm outperforms GA or SAA-GA, approaching the best solution within fewer simulations and has a constantly better RelOC within the simulation budget. In fact, in the deterministic SNPP case, MCLB-KG finds the true optimum within 80 100 iterations in most sample paths, while GA often stays in local optimum with RelOC values above 0.1 after reaching the 100 sampling budget. Figure 4.2 (d) shows how many times the alternatives in each region j ($j = 1, ..., 10$) are measured by each algorithm in three typical sample paths. As can be seen, GA tends to spend most time around certain area (near local optimum) while the MCLB-KG algorithm explores across the decision space more evenly. In fact, The global exploration of the MCLB-KG algorithm happens in earlier iterations accounting for larger uncertainties in the hyper-parameters and then the algorithm quickly identifies promising regions to spend more iterations in later sampling stages. In the stochastic SNPP setting, our algorithm also explores across the decision space while SAA-GA's searching is much more localized, similar to the observation in the deterministic case.

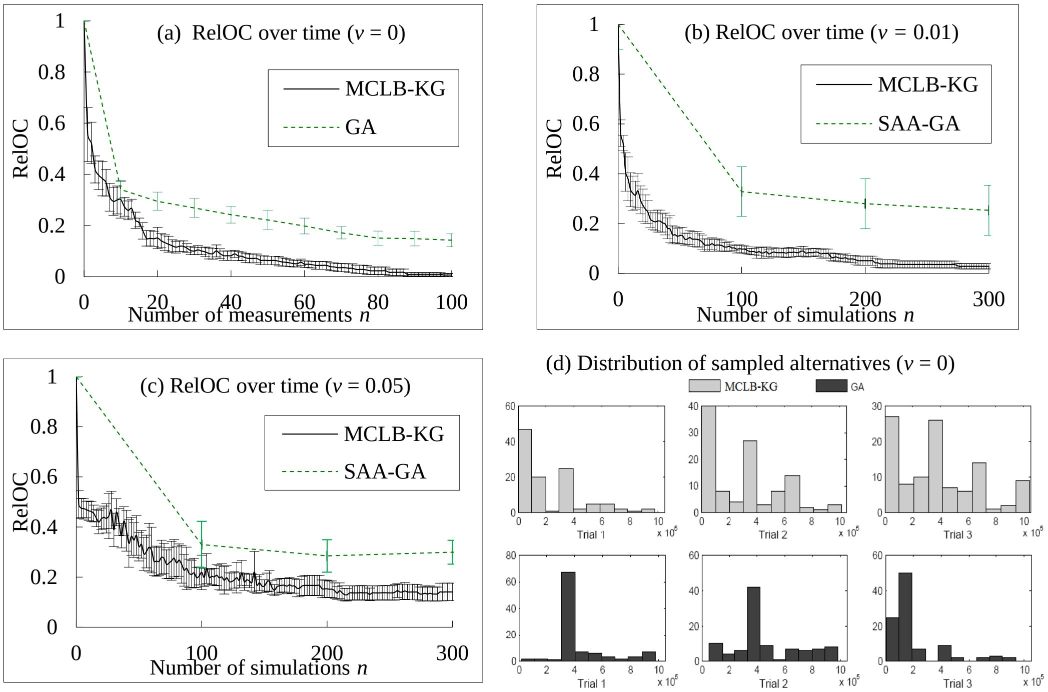
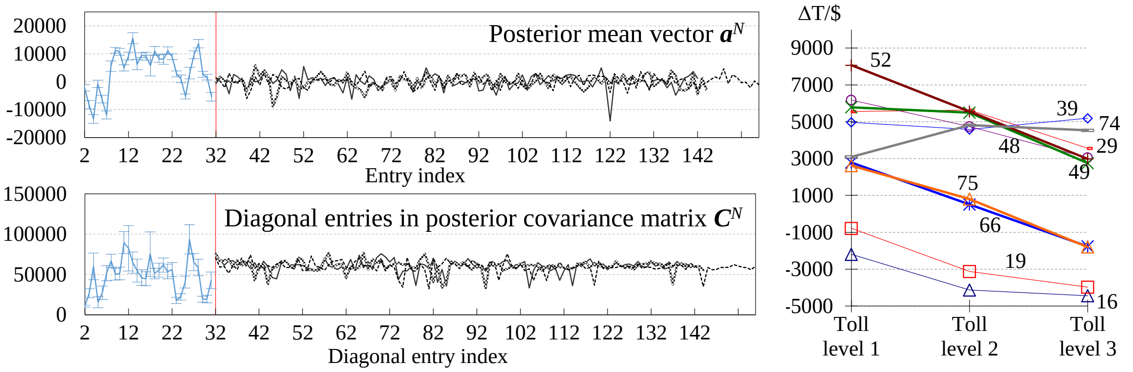


Figure 4.2: Performance comparison between two algorithms.

Take the case where $v = 0.01$ as an example, Figure 4.3(a) shows the entries of posterior mean vector a^N and diagonal entries of covariance matrix C^N resulting from MCLB-KG. We can see that the absolute values of the posterior means of model coefficients for most attributes are well above those of the sampled deviation terms, and the posterior variances of the deviation terms are smaller than those for the coefficients of toll attributes. This explains why the non-perfect additive linear models are useful for SNPP. These are also observed for the case where $v = 0.05$ (although the absolute values of a^N entries decrease) and the deterministic case. Under larger $v = 0.05$, the relative ranking among the posterior means of different model coefficients remain almost unchanged, and posterior variances of the toll attribute effects increase, which is not surprising. Based on these posterior means and $e = 2\$$, we compute and plot the marginal effects of toll rates for each preselected tollable link, as shown in Figure 4.3(b). We can see there are notable variations of the marginal effects across links and toll levels, which justifies that the belief model we used (i.e., using a separate coefficient η_i for each toll level at a specific link). The results also suggest that most links have positive expected marginal effect on travel time reduction at all toll levels, but interestingly, the expected marginal effect of link 66 and 75 are positive at toll level 1 and 2 but negative at toll level 3. Link 16 and 19 have negative expected marginal effect at all toll levels, indicating that the initial congestion level may not always be a good criteria for selecting candidate links.



(a) Posteriors on model coefficients (the lines on the left: posteriors of η_i's with ~95% CI, the lines on the right: posterior means of ζ_x's under 3 typical trials)

(b) Marginal effect at three toll levels (the link number is beside the line)

Figure 4.3: Posterior distributions on hyperparameters and marginal effects of toll attributes ($v = 0.01$).

We also note in the test that measurement decisions $x^{KG,n}$ are usually from the set $\{t_1, ..., t_{K_0}\}$ by MC sampling, which is a bigger set in earlier iterations ($n \leq 20$). However, in later iterations (after enough observations that make the belief upon those hyper-parameters relatively stable and outweigh the effect of non-informative priors), the MC step often selects only one alternative t_1 (i.e., $K_0 = 1$). Interestingly, this t_1 then often stalls for several iterations before a change is invoked by a relatively significant refinement of the belief in the linear model coefficients after sampling a new "promising" alternative from the set $\{s_1, ..., s_{K_1}\}$. This shows the effectiveness of MC sampling as well as the necessity of including extra alternatives in the candidate set S in each iteration, particularly when the number of alternatives is large.

Finally, Table 4.1 shows the average per iteration computation time of each algorithm over 10 sample paths, the MCLB-KG spends most time on sampling decision (MC step included), which is almost 10^3 times as long as that spent by the sampling step of GA (or SAA-GA). This is mainly due to computing the KG-factor over a bigger choice set especially during the earlier stages when candidate alternatives $\{t_1, ..., t_{K_0}\}$ are more diversified with larger K_0. Besides the doubled simulation time per iteration (due to evaluating $T_0(\omega)$ in addition to $T_x(\omega)$), another significant difference of the stochastic case compared to the deterministic case is that the average time spent on the MC sampling step increased by ~30% under $v = 0.05$. This is because during earlier iterations more candidate alternatives are generated due to bigger uncertainty on the hyper-parameters. Such uncertainty decreases significantly as measurements accumulate, but with K_0 drops in a slower rate compared to that in the case where $v = 0$ or $v = 0.01$. However, although the total computational time by MCLB-KG is bigger in this test network, it considerably reduces the total number of simulations needed for reaching a satisfactory RelOC compared to the SAA-GA. This will bring us substantial time savings for large networks when each simulation takes hours even days, which is usually the case for SNPP in practice (e.g., [140]). Therefore, the proposed MCLB-KG Algorithm can be very promising in solving real SNPP on large networks.

Table 4.1: Average computational time per iteration

v	Algorithm	Simulation (s)	Sampling decision (s)	Update (s)
0	GA	17.1	0.13	< 0.01
	MCLB-KG	16.5	41.4(57.5)[a]	0.81
0.01	SAA-GA	33.6	0.07	< 0.01
	MCLB-KG	33.4	39.5(62.4)[a]	0.79
0.05	SAA-GA	34.0	0.05	< 0.01
	MCLB-KG	33.6	37.8(76.4)[a]	0.67

[a] Time spent on KG-factor computing (time spent on MC sampling)

4.5 Conclusion

We have proposed a Bayesian R&S model for the Second-best Network Pricing Problem (SNPP) choosing toll locations and rates simultaneously. The large number of alternatives, combinatorial nature and random demand make the problem challenging. We adopt a linear belief KG policy to solve the SNPP. As an extension of [106] to the linear belief setting, MC-sampling of the hyper-parameters is proposed to reduce the choice set. Experiment results on a SNPP with $4^1$0 alternatives show good performance of the method and its superiority to the SAA-GA benchmark. We believe this is a promising tool for real-world SNPP under limited sampling budget. The successful application of the parameterized belief model tailored to SNPP also sheds lights on the underlying features of the problem itself as well as other transport network planning problems with similar nature.

BIBLIOGRAPHY

[1] Glpk (gnu linear programming kit). https://www.gnu.org/software/glpk/, 2012.

[2] 2015 New York City bridge traffic volumes. http://www.nyc.gov/html/dot/downloads/pdf/nyc-bridge-traffic-report-2015.pdf, 2015.

[3] How to use the I-15 express lanes. http://511sd.com/fastrak511sd/how-to-use-the-I-15-Express-Lanes, 2016.

[4] North texas express lane project, 2016. http://www.ntetexpress.com, 2016.

[5] Hudson River - George Washington Bridge wind statistics. http://wind.willyweather.com/nj/bergen-county/hudson-river–george-washington-bridge.html, April 2017.

[6] Wikipedia-George Washington Bridge. https://en.wikipedia.org/wiki/George_Washington_Bridge, April 2017.

[7] Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman E. Ozdaglar. Informational braess' paradox: The effect of information on traffic congestion. *CoRR*, abs/1601.02039, 2016.

[8] Sabreena Anowar, Naveen Eluru, and Marianne Hatzopoulou. Quantifying the value of a clean ride: How far would you bicycle to avoid exposure to traffic-related air pollution? In *TRB 96th Annual Meeting Compendium of Papers*, number 17-3351. Transportation Research Board, Jan 2017.

[9] H. M. Abdul Aziz and Satish V. Ukkusuri. Integration of environmental objectives in a system optimal dynamic traffic assignment model. *Computer-Aided Civil and Infrastructure Engineering*, 27:494–511, 2012.

[10] H. Bar-Gera. Transportation network test problems. http://www.bgu.ac.il/ bargera/tntp, 2013.

[11] M. Baykal-Gürsoy, W. Xiao, and K. Ozbay. Modeling traffic flow interrupted by incidents. *European Journal of Operational Research*, 195(1):127–138, May 2009.

[12] T. Bellemans, B. De Schutter, and B. De Moor. Model predictive control for ramp metering of motorway traffic: A case study. *Control Engineering Practice*, 14(7):757 – 767, July 2006.

[13] Alberto Bemporad and Manfred Morari. Control of systems integrating logic, dynamics, and constraints. *Automatica*, 35(3):407 – 427, March 1999.

[14] Eran Ben-Elia, Roberta Di Pace, Gennaro N. Bifulco, and Yoram Shiftan. The impact of travel informations accuracy on route-choice. *Transportation Research Part C: Emerging Technologies*, 26:146 – 159, Januray 2013.

[15] Antonio Bento, Daniel Kaffine, Kevin Roth, and Matthew Zaragoza-Watkins. The effects of regulation in the presence of multiple unpriced externalities: Evidence from the transportation sector. *American Economic Journal: Economic Policy*, 6(3):1–29, August 2014.

[16] Pia Bergendorff, Donald W. Hearn, and Motakuri V. Ramana. Congestion toll pricing of traffic networks. In Panos M. Pardalos, Donald W. Hearn, and William W. Hager, editors, *Network Optimization*, pages 51–71, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg.

[17] D. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Nashua, NH, 3rd edition, 2005.

[18] C. Bialik. New Yorkers will pay $ 56 a month to trim a minute off their commute. https://fivethirtyeight.com/features/new-yorkers-will-pay-56-a-month-to-trim-a-minute-off -their-commute/, 2016.

[19] Alexander Y. Bigazzi and Miguel A. Figliozzi. Congestion and emissions mitigation: A comparison of capacity, demand, and vehicle based strategies. *Transportation Research Part D*, 17(7):538–547, October 2012.

[20] Alexander Y. Bigazzi and Miguel A. Figliozzi. Marginal costs of freeway traffic congestion with on-road pollution exposure externality. *Transportation Research Part A*, 57:12–24, November 2013.

[21] Alexander Y. Bigazzi, Miguel A. Figliozzi, and Kelly J. Clifton. Traffic congestion and air pollution exposure for motorists comparing exposure duration and intensity. *International Journal of Sustainable Transportation*, 9(7):443–456, 2015.

[22] F. Borrelli, A. Bemporad, and M. Morari. *Predictive Control for Linear and Hybrid Systems*. Cambridge, 2015.

[23] Sebastien Boyer, Sebastien Blandin, and Laura Wynter. Stability of transportation networks under adaptive routing policies. *Transportation Research Part B: Methodological*, 81:886 – 903, November 2015.

[24] Mark Burris, Scott Nelson, Pete Kelly, Partha Gupta, and Youngjae Cho. Willingness to pay for high-occupancy toll lanes. *Transportation Research Record: Journal of the Transportation Research Board*, 2297:47–55, 2012.

[25] Gabriel Campanario. How many HOV-lane cheaters are there, and how many get caught? https://www.seattletimes.com/seattle-news/transportation/how-many-hov-lane-cheaters-are-there-and-how-many-get-caught/, June 2017.

[26] Carlos Carrion and David Levinson. Value of travel time reliability: A review of current evidence. *Transportation Research Part A: Policy and Practice*, 46(4):720–741, May 2012.

[27] Xiqun (Michael) Chen, Zheng Zhu, Xiang He, and Lei Zhang. Surrogate-based optimization for solving a mixed integer network design problem. *Transportation Research Record: Journal of the Transportation Research Board*, 2497(13):124–134, January 2015.

[28] C. Chung. A review and advance of high-occupancy toll lanes toll schemes. *Journal of the Eastern Asia Society for Transportation Studies*, 10(1):240–259, 2013.

[29] Jessica Coria, Jorge Bonilla, Maria Grundstrm, and Hkan Pleijel. Air pollution dynamics and the need for temporally differentiated road pricing. *Transportation Research Part A*, 75(C):178–195, 2015.

[30] Mengying Cui and David M Levinson. The greenest path: Comparing the effects of internal and external costs of motor vehicle pollution on route choice and accessibility. *Working Paper*, 2016.

[31] Inc. CVX Research. Software for disciplined convex programming. http://cvxr.com/, 2012.

[32] Carlos F. Daganzo. The cell transmission model, part II: Network traffic. *Transportation Research Part B: Methodological*, 29(2):79 – 93, 1995.

[33] Noam David and H. Oliver Gao. Using cellular communication networks to detect air pollution. *Environmental Science and Technology*, 50(17):9442–9451, August 2016.

[34] A.J. de Jong. Quality of real-time travel time information. Master's thesis, University of Twente, September 2012.

[35] Kusum Deep, Krishna Pratap Singh, M.L. Kansal, and C. Mohan. A real coded genetic algorithm for solving integer and mixed integer optimization problems. *Applied Mathematics and Computation*, 212(2):505 – 518, 2009.

[36] Elena G. Dorogush and Alex A. Kurzhanskiy. Modeling toll lanes and dynamic pricing control. *CoRR*, abs/1505.00506, 2015.

[37] Vivek Dua and Efstratios N. Pistikopoulos. An algorithm for the solution of multiparametric mixed integer linear programming problems. *Annals of Operations Research*, 99(1):123–139, Dec 2000.

[38] Liisa Ecola and Thomas Light. Equity and congestion pricing – a review of the evidence. Technical report, The RAND Corporation, Santa Monica, CA, 2009.

[39] Joakim Ekström, Ida Kristoffersson, and Nils-Hassan Quttineh. Surrogate-based optimization of cordon toll levels in congested traffic networks. *Journal of Advanced Transportation*, 50(6):1008–1033, 2016.

[40] Joakim Ekström, Agachai Sumalee, and Hong K. Lo. Optimizing toll locations and levels using a mixed integer linear approximation approach. *Transportation Research Part B: Methodological*, 46(7):834 – 854, August 2012.

[41] Alan L. Erera, Carlos F. Daganzo, and David J. Lovell. The access-control problem on capacitated fifo networks with unique o-d paths is hard. *Operations Research*, 50(4):736–743, August 2002.

[42] U.S. DOT Federal Highway Administration (FHWA). Status of the nations highways, bridges and transit: 2008 conditions and performances. https://www.fhwa.dot.gov/policy/2008cpr/, 2008.

[43] Federal Highway Administration (FHWA). Priced managed lane guide. Technical Report FHWA-HOP-13-007, U.S. Department of Transportation, https:

//ops.fhwa.dot.gov/publications/fhwahop13007/fhwahop13007.pdf, 2012.

[44] Lisa Fleischer, Kamal Jain, and Mohammad Mahdian. Tolls for heterogeneous selfish users in multicommodity networks and generalized congestion games. In *Proceedings of 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04)*, October 2004.

[45] Peter I. Frazier, Warren B. Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing*, 21:599–613, 2009.

[46] Spencer M. Free and James W. Wilson. A mathematical contribution to structure-activity studies. *Journal of Medicinal Chemistry*, 7(4):395–399, July 1964.

[47] Lina Fu and Rakesh Kulkarni. Model-based dynamic pricing algorithm for managed lanes. *Transportation Research Record: Journal of the Transportation Research Board*, 2333:74–79, 2013.

[48] United States Government Accountability Office (GAO). Road pricing can help reduce congestion, but equity concerns may grow. Technical Report GAO-12-119, Washington, DC, January 2012.

[49] Lauren M. Gardner, Hillel Bar-Gera, and Stephen D. Boyles. Development and comparison of choice models and tolling schemes for high-occupancy/toll (HOT) facilities. *Transportation Research Part B: Methodological*, 55:142 – 153, September 2013.

[50] Lauren M. Gardner, Avinash Unnikrishnan, and S. Travis Waller. Solution methods for robust pricing of transportation networks under uncertain demand. *Transportation Research Part C: Emerging Technologies*, 18(5):656 – 667, October 2010. Applications of Advanced Technologies in Transportation: Selected papers from the 10th AATT Conference.

[51] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Florida, 2004.

[52] Moshe Givoni. Re-assessing the results of the london congestion charging scheme. *Urban Studies*, 49(5):1089–1105, April 2012.

[53] Gabriel Gomes, Roberto Horowitz, Alex A. Kurzhanskiy, Pravin Varaiya,

and Jaimyoung Kwon. Behavior of the cell transmission model and effectiveness of ramp metering. *Transportation Research Part C: Emerging Technologies*, 16(4):485 – 513, August 2008.

[54] Harold Greenberg. An analysis of traffic flow. *Operations Research*, 7(1):79–85, February 1959.

[55] N. Groot, B. De Schutter, and H. Hellendoorn. Integrated model predictive traffic and emission control using a piecewise-affine approach. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):587–598, June 2013.

[56] Ren-Yong Guo, Hai Yang, Hai-Jun Huang, and Zhijia Tan. Link-based day-to-day network traffic dynamics and equilibria. *Transportation Research Part B: Methodological*, 71:248 – 260, January 2015.

[57] Ren-Yong Guo, Hai Yang, Hai-Jun Huang, and Zhijia Tan. Day-to-day flow dynamics and congestion control. *Transportation Science*, 50(3):982–997, August 2016.

[58] H.-M. Gutmann. A radial basis function method for global optimization. *Journal of Global Optimization*, 19(3):201–227, Mar 2001.

[59] J. Hale. *Functional Differential Equations*. Springer, 1971.

[60] Randy Halvorson and Kenneth R. Buckeye. High-occupancy toll lane innovations: I-394 MnPASS. *Public Works Management & Policy*, 10(3):242–255, January 2006.

[61] Donald W. Hearn and Mehmet B. Yildirim. *A Toll Pricing Framework for Traffic Assignment Problems with Elastic Demand*, pages 135–145. Springer US, Boston, MA, 2002.

[62] Andreas Hegyi, Bart De Schutter, and Hans Hellendoorn. Model predictive control for optimal coordination of ramp metering and variable speed limits. *Transportation Research Part C: Emerging Technologies*, 13(3):185 – 209, June 2005.

[63] R. Hemmecke, M. Köppe, J. Lee, and R. Weismantel. *50 Years of Integer Programming 1958-2008*, chapter 15 ”Nonlinear integer programming”. Springer-Verlag, Berlin, Germany, 2009.

[64] Winnie Hu. Congestion pricing falters in New York, again.

https://www.nytimes.com/2018/03/31/nyregion/congestion-pricing-new-york.html, March 2018.

[65] Rouba Ibrahim and Ward Whitt. Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research*, 59(5):1106–1118, 2011.

[66] O. İlker Kolak, Orhan Feyzioğlu, Ş. lker Birbil, Nilay Noyan, and Semih Yalçindağ. Using emission functions in modeling environmentally sustainable traffic assignment policies. *Journal of Industrial and Management Optimization (JIMO)*, 9(2):341–363, April 2013.

[67] Daniel J. Jacob. *Introduction to atmospheric chemistry*. Princeton University Press, 1999.

[68] Rajat Jain and J.Macgregor Smith. Modeling vehicular traffic flow using M/G/C/C state dependent queueing models. *Transportation Science*, 31(4):324–336, 11 1997.

[69] Rong-Chang Jou and Yi-Chun Yeh. Freeway passenger car drivers' travel choice behaviour in a distance-based toll system. *Transport Policy*, 27:11 – 19, May 2013.

[70] Seong Hee Kim and Barry L. Nelson. Recent advances in ranking and selection. In *Proceedings of the 2007 Winter Simulation Conference, WSC*, pages 162–172, 12 2007.

[71] Hideo Konishi. Uniqueness of user equilibrium in transportation networks with heterogeneous commuters. *Transportation Science*, 38(3):315–330, August 2004.

[72] Prashant Kumar, Lidia Morawska, Wolfram Birmili, Pauli Paasonen, Min Hu, Markku Kulmala, Roy M. Harrison, Leslie Norford, and Rex Britter. Ultrafine particles in cities. *Environment International*, 66:1 – 10, May 2014.

[73] Siriphong Lawphongpanich and Donald W. Hearn. An mpec approach to second-best toll pricing. *Mathematical Programming*, 101(1):33–55, Sep 2004.

[74] David Levinson. The value of advanced traveler information systems for route choice. *Transportation Research Part C*, 11(1):75–87, Feb 2003.

[75] David Levinson. Equity effects of road pricing: A review. *Transport Reviews*, 30(1):33–57, 2010.

[76] Jonathan I Levy, Jonathan J Buonocore, and Katherian von Stackelberg. Evaluation of the public health impacts of traffic congestion: a health risk assessment. *Environment Health*, 2010.

[77] Zhi-Chun Li, William H. K. Lam, S. C. Wong, and A. Sumalee. Environmentally sustainable toll design for congested road networks with uncertain demand. *International Journal of Sustainable Transportation*, 6(3):127–155, 2012.

[78] Bai Lihui, Hearn Donald W., and Lawphongpanich Siriphong. Decomposition techniques for the minimum toll revenue problem. *Networks*, 44(2):142–150, July.

[79] Weihua Lin, Amit Kulkarnim, and Pitu Mirchandani. Short-term arterial travel time prediction for advanced traveler information systems. *Intelligent Transportation Systems*, 8(3):143–154, 2004.

[80] Robin Lindsey, Terry Daniel, Eyran Gisches, and Amnon Rapoport. Pretrip information and route-choice decisions with stochastic travel conditions: Theory. *Transportation Research Part B*, 67:187–207, September 2014.

[81] X. Liu, G. Zhang, Y. Wu, and Y. Wang. Analyzing system performance for Washington state route 167. In *Transportation Research Board 89th Annual Meeting, Compendium of Papers DVD*, number 102816. Transportation Research Board, January 2010.

[82] Yingyan Lou, Yafeng Yin, and Jorge A. Laval. Optimal dynamic pricing strategies for high-occupancy/toll lanes. *Transportation Research Part C: Emerging Technologies*, 19(1):64 – 74, February 2011.

[83] X. Y. Lu, P. Varaiya, R. Horowitz, D. Su, and S. E. Shladover. A new approach for combined freeway variable speed limits and coordinated ramp metering. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 491–498, September 2010.

[84] Zhi-Quan Luo, Wing-Kin Ma, Anthony Man-Cho So, Yinyu Ye, and Shuzhong Zhang. Semidefinite relaxation of quadratic optimization problems. *IEEE Signal Processing Magazine*, 27(3):20–34, May 2010.

[85] Hani Mahmassani, Tian Hou, and Meead Saberi. Connecting network-wide travel time reliability and the network fundamental diagram of traffic flow. *Transportation Research Record: Journal of the Transportation Research Board*, 2391(8):80–91, January 2013.

[86] Hani Mahmassani, Tian Hou, and Meead Saberi. Connecting network-wide travel time reliability and the network fundamental diagram of traffic flow. *Transportation Research Record: Journal of the Transportation Research Board*, 2391:80–91, January 2013.

[87] Dimitra Michalaka, Yafeng Yin, and David Hale. Simulating high-occupancy toll lane operations. *Transportation Research Record: Journal of the Transportation Research Board*, 2396:124–132, 2013.

[88] MaïtéStéphan Mickael Beaud, Thierry Blayac. The impact of travel time variability and travelers risk attitudes on the values of time and reliability. *Transportation Research Part B*, 93(A):207–224, November 2016.

[89] S. Moghaddam and B. Hellinga. Real-time prediction of arterial roadway travel times using data collected by bluetooth detectors. *Transportation Research Record: Journal of the Transportation Research Board*, 2442:117–128, 2014.

[90] Juliane Müller, Christine A. Shoemaker, and Robert Piché. SO-I: a surrogate model algorithm for expensive nonlinear integer programming problems including global optimization applications. *Journal of Global Optimization*, 59(4):865–889, Aug 2014.

[91] L. Munoz, Xiaotian Sun, R. Horowitz, and L. Alvarez. Traffic density estimation with the cell transmission model. In *Proceedings of the 2003 American Control Conference, 2003.*, volume 5, pages 3750–3755, June 2003.

[92] Kai Nagel and Michael Schreckenberg. A cellular automaton model for freeway traffic. *Journal de Physique I*, 2(12):2221–2229, December 1992.

[93] A. Nagurney. A multiclass, multicriteria traffic network equilibrium model. *Mathematical and Computer Modelling*, 32(3-4):393–411, August 2000.

[94] Diana M. Negoescu, Peter I. Frazier, and Warren B. Powell. The knowledge-gradient algorithm for sequencing experiments in drug discovery. *INFORMS Journal on Computing*, 23(3):346–363, 2011.

[95] Matthew Neidell. Public information and avoidance behavior: Do people respond to smog alerts? Columbia University, Feb 2006.

[96] NYDEC. New York state department of environmental conservation (nydec) - air monitoring website. http://www.nyaqinow.net, 2017. Accessed: April 2017.

[97] Bureau of Public Roads. *Traffic Assignment Manual.* U.S. Department of Commerce, Urban Planning Division, Washington D.C., 1964.

[98] Jamol Pender, Richard H. Rand, and Elizabeth Wesson. Queues with choice via delay differential equations. *International Journal of Bifurcation and Chaos*, 2016.

[99] Jamol Pender, Richard H Rand, and Elizabeth Wesson. Strong approximations for queues with choice and constant delays. *Submitted for Publication*, 2017.

[100] Georgios Piliouras, Evdokia Nikolova, and Jeff S. Shamma. Risk sensitivity of price of anarchy under uncertainty. *ACM Transactions on Economics and Computation (TEAC)*, 5(1):1–27, November 2016.

[101] M. Powell. *The theory of radial basis function approximation in 1990, advances in numerical analysis, vol. 2: wavelets, subdivision algorithms and radial basis functions.* Oxford University Press, Oxford, 1992.

[102] W. B. Powell and I. O. Ryzhov. *Optimal Learning*. John Wiley and Sons, Hoboken, New Jersey, 2012.

[103] Ammon Rapoport, Eyran J. Gisches, Terry Daniel, and Robin Lindsey. Pre-trip information and route-choice decisions with stochastic travel conditions: Experiment. *Transportation Research Part B*, 68:154–172, October 2014.

[104] Paul I. Richards. Shock waves on the highway. *Operations Research*, 4(1):42–51, February 1956.

[105] O.M. Rouhani and H. Oliver Gao. An advanced traveler general information system for Fresno, California. *Transportation Research Part A*, 67:245–267, September 2014.

[106] I. O. Ryzhov and W. Powell. A monte carlo knowledge gradient method

for learning abatement potential of emissions reduction technologies. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pages 1492–1502, Dec 2009.

[107] Soodeh Saberian. Behavioral impacts of air quality alerts: Cycling and ozone alerts in sydney. *Working Paper*, November 2014. Available online: http://www.cireqmontreal.com/wp-content/uploads/2015/07/saberian.pdf.

[108] D. Schrank, B. Eisele, T. Lomax, and J. Bak. 2015 Urban Mobility Scorecard. Technical report, The Texas A&M Transportation Institute and INRIX, Inc., Aug 2015.

[109] Y. Sheffi. *Urban Transportation Newtorks: Equilibrium Analysis with Mathematical Programming Methods*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1985.

[110] S. P. Shepherd, A. D. May, D.S. Milne, and A. Sumalee. Practical algorithms for finding the optimal road pricing location and charges. In *European Transport Conference*, Cambridge, 2001.

[111] Simon Shepherd and Agachai Sumalee. A genetic algorithm based approach to optimal toll level and location problems. *Networks and Spatial Economics*, 4(2):161–179, Jun 2004.

[112] Mohammadali Shirazi and Hedayat Z. Aashtiani. Solving the minimum toll revenue problem in real transportation networks. *Optimization Letters*, 9(6):1187–1197, August 2015.

[113] H. Smith. *An Introduction to Delay Differential Equations with Applications to the Life Sciences*. Springer Science & Business Media, 2010.

[114] M.J. Smith. Existence, uniqueness, and stability of traffic equilibria. *Transportation Research*, 13B:295–304, February 1979.

[115] M.J. Smith, R. Liu, and R. Mounce. Traffic control and route choice: Capacity maximisation and stability. *Transportation Research Part B: Methodological*, 81:863 – 885, November 2015.

[116] S. Strogatz. *Nonlinear Dynamics and Chaos: with Applications to Physics, Biology, Chemistry, and Engineering*. Westview Press, 1994.

[117] Agachai Sumalee and Wei Xu. First-best marginal cost toll for a traffic network with stochastic demand. *Transportation Research Part B: Methodological*, 45(1):41 – 59, 2011.

[118] J. Thai and A. M. Bayen. State estimation for polyhedral hybrid systems and applications to the godunov scheme for highway traffic estimation. *IEEE Transactions on Automatic Control*, 60(2):311–326, February 2015.

[119] Jérôme Thai, Nicolas Laurent-Brouty, and Alexandre M. Bayen. Negative externalities of gps-enabled routing applications: A game theoretical approach. In *Proceedings of IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–4, Windsor Oceanico Hotel, Rio de Janeiro, Brazil, November 2016.

[120] Tomer Toledo, Omar Mansour, and Jack Haddad. Simulation-based optimization of hot lane tolls. *Transportation Research Procedia*, 6:189 – 197, 2015.

[121] Kenneth Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2002.

[122] E. T. Verhoef. Second-best congestion pricing in general static transportation networks with elastic demands. *Regional Science and Urban Economics*, 32(3):281–310, 2001.

[123] Eleni I. Vlahogianni, Matthew G. Karlafits, and John C. Golias. Short-term traffic forecasting: Where we are and where were going. *Transportation Research Part C*, 43(1):3–19, June 2014.

[124] Y. Wang and G. Zhang. A self-adaptive toll rate algorithm for high occupancy toll (hot) lane operations. Technical Report TNW200909, University of Washington and the U.S. Department of Transportation, 2009.

[125] Joseph White. Traffic jams cost U.S. drivers $1,200 a year: Study. https://www.usnews.com/news/us/articles/2017-02-20/traffic-jams-cost-us-drivers-1-200-a-year-study, 2017.

[126] Ward Whitt. Improving service by informing customers about anticipated delays. *Management science*, 45(2):192–207, 1999.

[127] WHO. Ambient air pollution: A global assessment of exposure and bur-

den of disease. Technical report, World Health Organization (WHO), February 2016.

[128] Wen-Xiang Wua and Hai-Jun Huang. Finding anonymous tolls to realize target flow pattern in networks with continuously distributed value of time. *Transportation Research Part B*, 65:31–46, July 2014.

[129] Hai Yang and Hai-Jun Huang. Principle of marginal-cost pricing: how does it work in a general road network? *Transportation Research Part A: Policy and Practice*, 32(1):45 – 54, 1998.

[130] Hai Yang and Hai-Jun Huang. The multi-class, multi-criteria traffic network equilibrium and systems optimum problem. *Transportation Research Part B*, 38(1):1–15, January 2004.

[131] Hai Yang and Xiaoning Zhang. Optimal toll design in second-best link-based congestion pricing. *Transportation Research Record: Journal of the Transportation Research Board*, 1857:85–92, 2003.

[132] Li Yang, Romesh Saigal, and Hao Zhou. Distance-based dynamic pricing strategy for managed toll lanes. *Transportation Research Record: Journal of the Transportation Research Board*, 2283:90–99, 2012.

[133] W. Yi, K. Lo, T. Mak, K. Leung, Y. Leung, and M. Meng. A survey of wireless sensor network based air pollution monitoring systems. *Sensors*, 15(12):31392–31427, December 2015.

[134] Yafeng Yin and Yingyan Lou. Dynamic tolling strategies for managed lanes. *Journal of Transportation Engineering*, 135(2):45–52, 2009.

[135] Ding Zhang and Anna Nagurney. On the local and global stability of a travel route choice adjustment process. *Transportation Research Part B: Methodological*, 30(4):245 – 262, August 1996.

[136] Kai Zhang and Stuart Batterman. Air pollution and health risks due to vehicle traffic. *Science of Total Environment*, 450-451:307–316, April 2013.

[137] X. Zhang, X.B. Zhang, and X. Chen. Valuing air quality using happiness data: The case of China. *Ecological Economics*, 137:29 – 36, 2017.

[138] Feng Zhu and Satish V. Ukkusuri. A reinforcement learning approach for

distance-based dynamic tolling in the stochastic network environment. *Journal of Advanced Transportation*, 49(2):247–266, June 2014.

[139] Y. Zhu, T. Smith, M. Davis, J. Levy, R. Herrick, and H. Jiang. Comparing gravimetric and real-time sampling of pm2.5 concentrations inside truck cabins. *Journal of Occupational and Environmental Hygiene*, 8(11):662–672, November 2011.

[140] Zhi Zuo, Ryo Kanamori, Tomio Miwa, and Takayuki Morikawa. *Comparison of Cordon and Optimal Toll Points Road Pricing Using Genetic Algorithm*, pages 535–544.