http://dx.doi.org/10.1016/j.trb.2019.11.002

A dynamic user equilibrium model for multi-region macroscopic fundamental diagram systems with time-varying delays

Y. P. Huang [a,b], J. H. Xiong [a], A. Sumalee [b], N. Zheng [c], W.H.K. Lam [b], Z.B. He [d], R. X. Zhong [a,b,*]

[a] *Guangdong Key Laboratory of Intelligent Transportation Systems, School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou, China*

[b] *Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China*

[c] *Institute of Transport Studies, Department of Civil Engineering, Monash University, Australia*

[d] *College of Metropolitan Transportation, Beijing University of Technology, Beijing, China*

Abstract

Macroscopic fundamental diagram (MFD) has been widely used for aggregate modeling of urban traffic network dynamics to tackle the dimensionality problem of microscopic approaches. This paper contributes to the state-of-the-art by proposing a dynamic user equilibrium (DUE) model that enables simultaneous route choice and departure time choice under the MFD framework for various applications such as park-and-ride, vehicle dispatching and relocation. To better capture the traffic flow propagation and to adapt to the fast time-varying demand, the state-dependent travel time function is integrated into the MFD dynamics as an endogenous time-varying delay. The multi-region MFD dynamics with saturated state and inflow constraints is then used as the network loading model to formulate the DUE model through the lens of the differential variational inequality. Necessary conditions for the DUE are analytically derived using the Pontryagin minimum principle. Difficulties raised in handling the dynamic state-dependent nonlinear travel time functions, state and inflow constraints are addressed without model linearization nor enforcing constant delay assumption as conventionally done in the literature. The additional cost induced by inflow capacity and accumulation constraints can capture the hypercongestion represented by the downward sloping part of the MFD without actually activating traffic congestion. Numerical examples solved by using time-discretization solution algorithm illustrate the DUE characteristics and the corresponding dynamic external costs induced by constraints.

Keywords: Macroscopic fundamental diagram, endogenous time-varying delay, simultaneous route choice and departure time choice, dynamic user equilibrium, saturated state and input constraints.

1. Introduction

Macroscopic fundamental diagram (MFD), establishing a mapping from the network accumulation to the trip completion rate, has been widely used for aggregate modeling of urban traffic network dynamics

*Corresponding author

E-mail addresses: huangyp945@gmail.com (Y. P. Huang)., xiongjh3@mail2.sysu.edu.cn (J. H. Xiong)., asumalee@gmail.com (A. Sumalee)., Nan.Zheng@monash.edu (N. Zheng)., cehklam@polyu.edu.hk (W.H.K. Lam)., he.zb@hotmail.com (Z.B. He)., zhrenxin@mail.sysu.edu.cn (R. X. Zhong).

under stationary traffic assumption, i.e., static or slow-varying demand while the traffic is homogeneously distributed (Daganzo and Geroliminis, 2008; Buisson and Ladier, 2009; Geroliminis and Sun, 2011; Saberi et al., 2014; Laval and Castrillón, 2015). Using the MFD framework, a heterogeneous city can be partitioned into multiple homogeneous regions with each represented by a well-defined MFD model. The MFD framework has been applied to modeling and management of urban traffic systems including route guidance management (Yildirimoglu et al., 2015; Knoop et al., 2012), traffic control in urban networks (Geroliminis et al., 2013; Keyvan-Ekbatani et al., 2012, 2015a,b; Haddad et al., 2013; Haddad, 2017a,b; Zhong et al., 2018a) and parking management (Zheng and Geroliminis, 2016). Using the MFD framework to describe the dynamics of large-scale urban traffic network, Ramezani and Nourinejad (2018) developed a dynamic bimodal (cars and taxis) traffic modeling and control strategy, i.e., taxi dispatching, to improve urban mobility and to mitigate congestion in cities based on the Model Predictive Control (MPC). Compared with the existing vehicle dispatching systems, the MFD framework is of analytically tractable nature (Ramezani and Nourinejad, 2018) and gives rise to a promising solution to the challenge of spatial dimensionality in the meanwhile. For vehicle dispatching and relocation, it is necessary to enable route choice and departure time choice.

Recent efforts have been dedicated to combining the dynamic traffic assignment (DTA) and the MFD framework by incorporating route choices or preferences for networks modeled by multiple MFD regions. Drivers aim to decrease their travel times by choosing a different sequence of regions to their destinations (Knoop et al., 2012; Haddad et al., 2013; Yildirimoglu and Geroliminis, 2014; Yildirimoglu et al., 2015; Ampountolas et al., 2017). For mixed network with urban regions and freeways, Haddad et al. (2013) developed a dynamic simple route choice model to account for the change in route choice decision in response to the new control wherein travelers choose the route with minimum instantaneous travel time. Knoop et al. (2012); Yildirimoglu and Geroliminis (2014) integrated the MFD dynamics with aggregated route choice dynamics to develop a region-based route choice model using hierarchical control framework. Approximate user equilibrium (UE) and system optimum (SO) conditions for route choice behaviors under the MFD framework were further pursued by Yildirimoglu et al. (2015). Recently, Ampountolas et al. (2017) adopted the C-Logit model to describe the route choice behavior to simulate drivers' adaptiveness to traffic conditions with the presence of a robust perimeter and boundary flow controller. Combining Vickrey's bottleneck model with the MFD (that describes the network exit function), Amirgholy and Gao (2017) developed a bathtub model for approximating the user equilibrium of the morning commute problem under several assumptions, e.g., the equilibrium trajectory can be well approximated by a quadratic function. Yildirimoglu et al. (2018) built a two-level route guidance system based on the MPC scheme to minimize the total travel time. However, these existing methods are either algorithmic or numerical (Laval et al., 2018). Laval et al. (2018) considered the delay in the DUE model analytically for a simple network with a freeway as an alternative to the city street network modeled by the MFD dynamics wherein several definitional constraints on the state and input are ignored. As reviewed in Aghamohammadi and Laval (2018), neither departure time choice modeling nor capacity constraint has been well incorporated in the MFD literature. Route choice/guidance in conjunction with the departure time choice for general traffic network under the MFD framework is still unexplored in the MFD literature. This paper fills these gaps by proposing a DUE model that considers route choice and departure time choice simultaneously for the MFD framework under capacity constraints.

In the literature, the MFD dynamics regarding the vehicle accumulation is postulated in the ordinary differential equation formulation. As claimed in Mariotte et al. (2017), inconsistent phenomenon such as outflow is reacting too fast with respect to the rapid change in the inflow can be observed. However, the propagation speed of this information should never be faster than the free-flow speed. To overcome this

limitation, travel time should be explicitly incorporated in the flow propagation. The regional travel time as a function of the network accumulation (thus state-dependent and endogenous) can be time-varying due to the dynamic traffic conditions such as congestion onset and dissolve. In the literature, there are two approaches for incorporating the travel time in the flow propagation under the MFD framework. The travel time function is regarded as a time delay in the control input when it is explicitly integrated in the MFD dynamics (Keyvan-Ekbatani et al., 2012; Aboudolas and Geroliminis, 2013; Haddad and Mirkin, 2016; Mirkin et al., 2016). Haddad and Zheng (2018) incorporated the time delay in the state dynamics to better capture the traffic flow propagation using the MFD dynamics to improve the adaptive perimeter control strategies. The other approach is termed as the trip-based model (Arnott, 2013; Fosgerau, 2015; Arnott et al., 2016; Lamotte and Geroliminis, 2018; Mariotte et al., 2017; Arnott and Buli, 2018; Mariotte and Leclercq, 2019), wherein the travel time is assumed to depend on the time-varying traffic speed throughout the trip duration. In this paper, we stick to the first approach considering several technical barriers of the trip-based formulation. First, as claimed in Arnott and Buli (2018), the mathematical tool for tackling the trip-based model is not mature even for the basic single-region trip-based model wherein the trip-length is homogeneous. General analytical properties of the trip-based models are very difficult (somehow "intractable" by quoting the key message from Arnott and Buli, 2018). Due to such difficulties in dealing with the single-region trip-based model, the progress of extending the single region trip-based model to the multi-region case is slow while it has not yet reached a consensus in the literature. Recent progress on the multi-region trip-based model by Mariotte and Leclercq (2019) focuses on the numerical treatment rather than the analytical property. Therefore, we do not use the trip-based formulation in pursuing the analytical properties of the DUE with simultaneous route and departure time choices as well as capacity constraints. To simplify the control design problem, the delays (and thus the regional travel time functions) are assumed to be constant in the traffic flow dynamics in Keyvan-Ekbatani et al. (2012); Aboudolas and Geroliminis (2013); Haddad and Mirkin (2016); Mirkin et al. (2016); Haddad and Zheng (2018). However, due to the route and departure time choices, travel demand can be fast time-varying and network traffic state varies with respect to congestion. The state-dependent dynamic delay would be better to adapt to this time-varying nature of travel demand and network traffic state. But, how to handle such kind of endogenous time-varying delay is still missing in transportation literature, which is marked as a future research topic by Haddad and Zheng (2018). On the other hand, linearization technique is often adopted in the MFD literature, see e.g., Keyvan-Ekbatani et al. (2012); Aboudolas and Geroliminis (2013); Haddad and Mirkin (2016); Mirkin et al. (2016); Haddad and Zheng (2018). As discussed in Zhong et al. (2018a,b), linearization can only look into the local properties around the reference point to which the linearization is performed. However, evaluation of the region of attraction of the equilibrium of the linearized MFD system is difficult. To avoid the hurdle caused by linearization and constant delay assumption, we stick to the nonlinear MFD dynamics with dynamic state-dependent delays and state and input saturated constraints.

Following the general mathematical tool for the dynamic user equilibrium problem with route choice and/or departure time choice, i.e., the differential variational inequality (DVI) (Friesz et al., 1993, 2001; Friesz and Han, 2018), this research proposes a DVI formulation of the DUE that simultaneously considers the route choice and departure time choice under the MFD framework. An urban traffic network is assumed to be partitioned into several regions with each region admits a well-defined MFD. The dynamics of each region is described by a well-calibrated MFD model with inherent time-varying delay in traffic flow propagation and saturated constraints. It is assumed that the drivers would choose their departure times and routes to minimize their generalized travel costs while the constraints on state and input are captured by certain inequality constraints. Optimal control is applied to derive the necessary conditions for the DUE analytically under the umbrella of the Pontryagin minimum principle. Moreover, numerical approximations

to the DUE are solved by constrained nonlinear optimization algorithm after time-discretization to verify the validation of the proposed optimal conditions.

Paper organization: Section 2 introduces the MFD model with time delay for network representation. To model simultaneous route choice and departure time choice behaviors for general MFD systems with time-varying delays, Section 3 formulates the DUE as a DVI problem and analyzes the necessary conditions of the DUE by reformulating it as an optimal control with inequality constraints on both state and control. Equilibrium conditions are derived analytically through the lens of the Pontryagin minimum principle. Section 4 devises a numerical algorithm to solve the proposed DUE problem. Several numerical examples are worked out to illustrate the properties of the DUE in Section 5. Finally, Section 6 concludes the paper. Companion materials and proofs are presented in the appendix.

2. Network representation and the MFD system with state-dependent delay

An urban traffic network is partitioned into M regions, labeled as $R_1, R_2, \cdots, R_j, \cdots, R_M$, on condition that the traffic is homogeneously distributed within each region and thus each admits a well-defined MFD. Under this assumption, for a specific region R, there exists a concave "Production-MFD" function establishing a mapping from network accumulation $n_R(t)$ (veh) to the network production $P_R(n_R(t))$ (veh.m/s). The network accumulation, i.e., $n_R(t)$, is the total number of vehicles in region R at time t. The network average space-mean speed V_R (m/s) per Edie's definition is evaluated as $V_R(n_R(t)) = \frac{P_R(n_R(t))}{n_R(t)}$. Assuming the average trip length in region R is L_R, the network output (or exit) function, also known as the MFD, can be deduced as $G_R(n_R) = \frac{P_R(n_R)}{L_R}$. It is assumed that $G_R(n_R)$ (veh/s) is a concave function of $n_R(t)$.

By definition, the average network travel time function of region R at time t is a function of the network accumulation $n_R(t)$ at time t, which is defined as

$$h_R(n_R(t)) = \frac{L_R}{V_R(n_R(t))} = \frac{L_R}{\frac{P_R(n_R(t))}{n_R(t)}} = \frac{n_R(t)}{\frac{P_R(n_R(t))}{L_R}} = \frac{n_R(t)}{G_R(n_R(t))} \tag{1}$$

Note that the free flow travel time of region R can be obtained by applying L'Hospital's rule as

$$h_R(n_R)\big|_{n_R\to 0} = \lim_{n_R\to 0} \frac{n_R}{G_R(n_R)} = \lim_{n_R\to 0} \frac{\dot{n}_R}{\dot{G}_R(n_R)},$$

where we have omitted t for simplicity. For a single-region MFD system, its dynamics evolves according to the flow conservation:

$$\frac{\mathrm{d}n_R(t)}{\mathrm{d}t} = q_R(t) - G_R(n_R(t))$$

where $q_R(t)$ denotes the inflow rate to region R at time t. Or equivalently,

$$n_R(t) = n_R(0) + \int_0^t \Big(q_R(s) - G_R(n_R(s))\Big)\,\mathrm{d}s \tag{2}$$

where $n_R(0)$ denotes the initial network accumulation of R.

As claimed in Mariotte et al. (2017), inconsistent phenomenon such as outflow is reacting too fast with respect to the rapid change in the inflow can be observed. However, the propagation speed of this information should never be faster than the free-flow speed. To overcome this limitation, travel time should be explicitly incorporated in the flow propagation. Following Arnott and Buli (2018), flow conservation regarding travel time function $h_R(n_R(t))$ for vehicles entering R at time t can be defined as

$$N_{out}\left(t + h_R(n_R(t))\right) = N_{in}(t), \text{ or } \int_0^t q_R(s)\,\mathrm{d}s = \int_0^{t+h_R(n_R(t))} G_R(n_R(s))\,\mathrm{d}s \tag{3}$$

According to this definition, network accumulation at time t can be expressed as

$$\begin{aligned} n_R(t) &= N_{in}(t) - N_{out}(t) = N_{out}(t + h_R(n_R(t)) - N_{out}(t) \\ &= \int_0^{t+h_R(n_R(t))} G_R(n_R(s))\,\mathrm{d}s - \int_0^t G_R(n_R(s))\,\mathrm{d}s = \int_t^{t+h_R(n_R(t))} G_R(n_R(s))\,\mathrm{d}s \end{aligned}$$

By differentiating (3), the relationship between the inflow and outflow rate is given as

$$\Big(1 + \dot{h}_R(n_R(t))\Big) G_R\Big(t + h_R(n_R(t))\Big) = q_R(t)$$

Since both $G_R(\cdot)$ and $q_R(\cdot)$ are nonnegative, $1 + \dot{h}_R(n_R(t)) \geq 0$ so as to make this nonnegative be fulfilled. In the DTA literature, this condition is known as First-In-First-Out principle. If $1 + \dot{h}_R(n_R(t)) > 0$, we have

$$G_R\Big(t + h_R(n_R(t))\Big) = \frac{q_R(t)}{1 + \dot{h}_R(n_R(t))}. \tag{4}$$

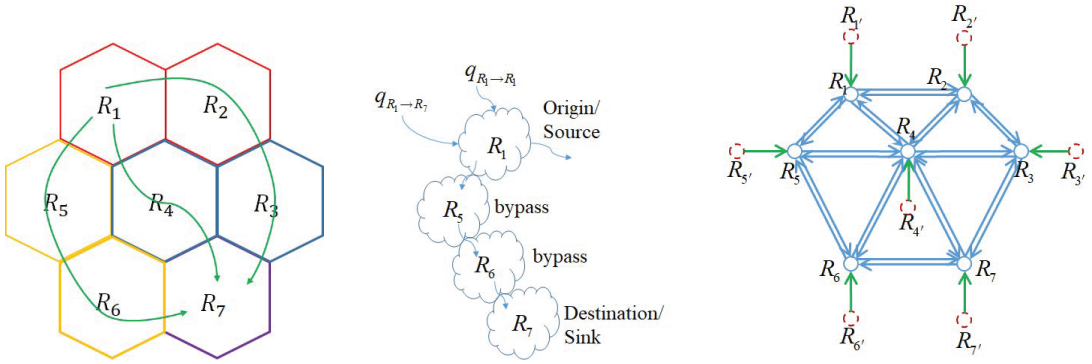
The FIFO assumption is enforced on the travel time function in the literature. In line with this, we also enforce the FIFO condition here to guarantee nonnegative flows.

Remark 2.1. The travel time function defined in (1) is a nonlinear function regarding the traffic state $n(t)$. Note that only linear travel-time functions can we be sure that the FIFO is well satisfied regardless of the inflow profile, see e.g., Friesz et al. (1993); Daganzo (1995); (Xu et al., 1999, Theorem 3.2, Corollary 3.1); (Zhu and Marcotte, 2000, Theorem 5.1); Carey and Mccartney (2002); Nie and Zhang (2005). However, the FIFO principle should not be the only criterion used in choosing travel-time functions as claimed in Carey et al. (2014) because of several critical drawbacks of the linear travel time function: a) linearity assumption is contradicted by empirical evidence from nonlinear speed-density and flow-density functions; b) linear travel-time function can substantially overestimate link travel time due to the double-counting effect (Nie and Zhang, 2005); c) linear travel-time functions cannot capture the congestion represented by the downward sloping part of the flow-density diagrams. Carey et al. (2014) thus proposed to employ nonlinear functions with treatment to avoid the FIFO violation. By imposing an upper bound to the link inflow profile and increasing the link traversal time of the follow-on vehicles (that violate FIFO), they ensured that FIFO holds. The remaining challenge is that, changing the exit time of vehicles violating FIFO would also change the analytical structure of the model. Our work follows closely this line. In this paper, we will use the approach of inflow capacity and accumulation constraints to enforce the FIFO condition for a nonlinear travel time function, see e.g., **DUE II**, the discussion in Remark 3.2 and the proof outlined in Appendix E. ∎

To enable route choice in the multi-region MFD system, an urban network is divided into a set of connected regions as depicted in Figure 1(a). Figure 1(b) shows an example of a path connecting the origin R_1 and the destination R_7. The path flow would bypass two regions R_5 and R_6 successively before reaching the destination which is denoted as

$$p = \{R_1, R_5, R_6, R_7\} \tag{5}$$

The network and its interconnection are described by a directed graph as depicted in Figure 1(c). For travel demand whose origin and destination are in the different regions, it is assumed that the flow is generated from a dummy source $R_{i'}$ and terminates in region R_j, $i \neq j$, where $R_{i'}$ denotes the dummy source and R_i the sink in region R_i. Separating the source and sink within the same region is important since a



(a) A schematic diagram of an urban traffic network divided into multiple regions

(b) Movements along different MFD regions and route choice behavior

(c) Directed graph representation of multi-region MFD system with route choice

Figure 1: Multi-region MFD system and the directed graph representation

trip can be started and ended in the same region with a travel time equal to the average travel time of the region under the MFD framework while this is prohibited in the conventional directed graph representation of traffic network.

By the directed graph representation, we can define sequential movements along different MFD regions using the concept of path/route by a sequence of connected regions, e.g.,

$$\mathcal{R}^p \triangleq \{R_{p,0}, R_{p,1}, R_{p,2}, \cdots, R_{p,i}, \cdots, R_{p,m(p)}\}, \ p \in \mathcal{P}$$

where $R_{p,0}$ refers to the dummy source of path p, i.e., the origin; $m(p)$ is the number of regions along path p; $\mathcal{P}$ denotes the set of all paths connecting the network. For example, (5) is reordered as

$$p = \{R_{p,0}, R_{p,1}, R_{p,2}, R_{p,3}, R_{p,4}\} \triangleq \{R_{1'}, R_1, R_5, R_6, R_7\}$$

To enable route choice and departure time choice for the MFD framework, for a specific path p, we adopt the concept of departure rate $q^p(t)$ to a particular time instant t in line with the DTA literature. Along the path, flow exiting from the preceding upstream region is taken as input to the downstream region. For example, $G^p_{R_{p,i-1}}(t)$, the flow exiting from upstream $R_{p,i-1}$ by path p is regarded as input to the downstream $R_{p,i}$ by path p. This flow propagation mechanism proceeds until this amount of flow arrives the destination region $R_{p,m(p)}$, then we have the following flow propagation dynamics for the regions along path p:

$$\frac{\mathrm{d}n^p_{R_{p,1}}(t)}{\mathrm{d}t} = q^p(t) - G^p_{R_{p,1}}(t), \quad \forall p \in \mathcal{P} \tag{6a}$$

$$\frac{\mathrm{d}n^p_{R_{p,i}}(t)}{\mathrm{d}t} = G^p_{R_{p,i-1}}(t) - G^p_{R_{p,i}}(t), \quad \forall p \in \mathcal{P}, \ i \in [2, m(p)] \tag{6b}$$

where $n^p_{R_{p,i}}(t)$ denotes the network accumulation state of $R_{p,i}$ contributed by path p at time t. This flow propagation law is inspired by that of the Whole Link Model (WLM) investigated in Friesz et al. (1993); Zhong et al. (2011, 2012). As aforementioned, each region admits a time delay/travel time function $h_{R_{p,i}}\left(n_{R_{p,i}}(t)\right)$, which defines the average travel time required to traverse region $R_{p,i}$, i.e.,

$$h_{R_{p,i}}\left(n_{R_{p,i}}(t)\right) = \frac{n_{R_{p,i}}(t)}{G_{R_{p,i}}(n_{R_{p,i}}(t))} \tag{7}$$

where $n_{R_{p,i}}(t)$ denotes the total accumulation state of the i^{th} region along path p at time t. We denote $\tau^p_{R_{p,i}}(t)$ the exit time from the i^{th} region $R_{p,i}$ along path p, which is also the entry time to $R_{p,i+1}$ of vehicles departing from the origin, i.e., $R_{p,0}$, of path p at time t.

$$\begin{aligned} \tau^p_{R_{p,1}}(t) &= t + h_{R_{p,1}}\left(n_{R_{p,1}}(t)\right), \ \forall p \in \mathcal{P} \\ \tau^p_{R_{p,i}}(t) &= \tau^p_{R_{p,i-1}}(t) + h_{R_{p,i}}\left(n_{R_{p,i}}(\tau^p_{R_{p,i-1}}(t))\right), \ \forall p \in \mathcal{P}, \ i \in [2, \ m(p)] \end{aligned} \tag{8}$$

To enable route choice, journey time of a path needs to be evaluated. The nested delay operator proposed by Friesz et al. (1993) is adopted for this purpose, see Figure 2:

$$h^p(t, \boldsymbol{n}) = \sum_{i=1}^{m(p)} h_{R_{p,i}}(\tau^p_{R_{p,i-1}}(t)) = \tau^p_{R_{p,m(p)}} - t, \ \forall p \in \mathcal{P}$$

where $\boldsymbol{n} = (n_{R_i} : \ \forall R_i \in \mathcal{R})$, and for convenience of presentation, $\tau^p_{R_{p,0}} = t$.

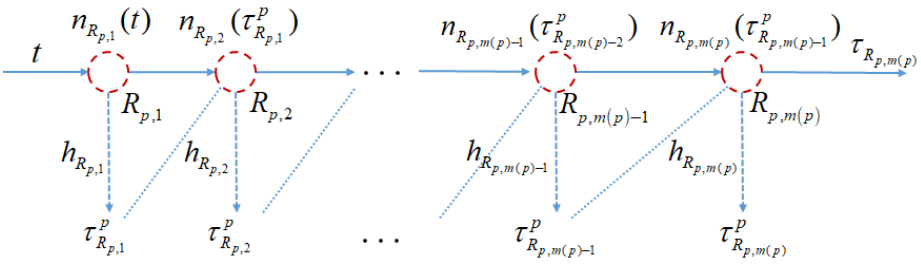


Figure 2: Journey time progression by nested delay operator for a path with $m(p)$ regions

To enforce departure time choice, an early/late arrival penalty, say $\kappa(\chi)$, is required. Combining this penalty with the journey time achieves the effective path delay/cost operators, i.e.,

$$\Psi^p(t, \boldsymbol{n}) = h^p(t, \boldsymbol{n}) + \kappa(\chi), \quad \forall p \in \mathcal{P} \tag{9}$$

where $h_p(t, \boldsymbol{n})$ is a function of accumulative state $\boldsymbol{n}$, and χ is the difference between actual and preferred arrival time t^*, $\chi = t + h^p(t, \boldsymbol{n}) - t^*$, with $t^* < T$. Note that given a traffic network described by the MFD system, the network accumulation is a direct result of the inflow to the region. Under the perimeter control framework, one can change the network accumulation by controlling the boundary flow to the region. To highlight the control variables, i.e., the path inflow rates $\boldsymbol{q}$, we write the path cost as

$$\Psi^p(t, \boldsymbol{q}) \triangleq \Psi^p(t, \boldsymbol{n}), \ \forall p \in \mathcal{P}$$

3. Analysis of dynamic user equilibrium

In this section, DUE is analyzed for general MFD systems with time-varying delays, inflow capacity and accumulation constraints considering simultaneous route choice and departure time choice. It is assumed in the DUE that each driver chooses a route and a departure time such that his/her own travel cost is minimized. The definition of DUE adopted here can be traced back to Friesz et al. (1993), and is also discussed in detail by Friesz et al. (2001); Friesz and Han (2018).

Definition 3.1. Dynamic user equilibrium (Friesz and Han, 2018). For any $\boldsymbol{q}^* \in \breve{\Lambda}$ with

$$\breve{\Lambda} = \left\{ \boldsymbol{q} : \boldsymbol{q}(t) \geq 0, \ \sum_{p \in \mathcal{P}_w} \int_0^T q^p(t)\, \mathrm{d}t = Q_w, \ \forall p \in \mathcal{P}, \ \forall t \in [0, T] \right\}, \tag{10}$$

the vector of path flows as $\boldsymbol{q} = (q^p : p \in \mathcal{P})$ and any nonnegative vector $\boldsymbol{\phi} = (\phi_w : w \in \mathcal{W}) \in R_+^{|W|}$, the pair $(\boldsymbol{q}^*, \boldsymbol{\phi})$ is a simultaneous departure-time-and-route-choice dynamic user equilibrium if and only if the following two conditions are satisfied for all $p \in \mathcal{P}_w$ and for all $w \in \mathcal{W}$:

$$\begin{array}{rcl} q^{p\,*}(t) > 0 & \Rightarrow & \Psi^p(t, \boldsymbol{q}^*) = \phi_w, \\ \Psi^p(t, \boldsymbol{q}^*) > \phi_w & \Rightarrow & q^{p\,*}(t) = 0, \end{array} \tag{11}$$

where ϕ_w is the smallest travel cost for the OD pair w given by

$$\phi_w = \min\{\varrho^p : p \in \mathcal{P}_w\} > 0, \ \forall w \in \mathcal{W}, \ \text{and } \varrho^p = \operatorname{ess\,inf}\{\Psi^p(t, \boldsymbol{q}) > 0 : t \in [0, \ T]\}, \ \forall p \in \mathcal{P}$$

where ess inf is essential infimum which defines the largest essential lower bound for a given function f wherein all $\inf f \leq \operatorname{ess\,inf} f$. ∎

For a set of connected regions outlined in Section 2, consider a finite time planning horizon $t \in [0, T]$ with $T > 0$, each origin-destination (OD) pair $w \in \mathcal{W}$ is with a fixed total amount of demand Q_w to be served. Similar to Friesz et al. (2001); Zhong et al. (2011), for a specific OD w, the DUE with simultaneous route choice and departure time choice can be formulated as the following differential variational inequality (DVI): find $(\boldsymbol{n}^*, \boldsymbol{q}^*) \in \Gamma$ such that

Problem 3.1. DUE I:

$$\left\langle \boldsymbol{\Psi}\left(t, \boldsymbol{q}^*\right), \left(\boldsymbol{q} - \boldsymbol{q}^*\right) \right\rangle = \sum_{p \in \mathcal{P}} \int_0^T \Psi^p\left(t, \boldsymbol{q}^*\right)\left(q^p(t) - q^{p\,*}(t)\right) \mathrm{d}t \geq 0, \quad \forall\, (\boldsymbol{n}, \boldsymbol{q}) \in \Gamma \tag{12}$$

where Γ satisfies the flow propagation, state and input constraints specified in (13a)-(13k).

$$\frac{\mathrm{d}n^p_{R_{p,1}}(t)}{\mathrm{d}t} = q^p(t) - G^p_{R_{p,1}}(t), \quad \left(\alpha^p_{R_{p,1}}\right) \quad \forall p \in \mathcal{P} \tag{13a}$$

$$\frac{\mathrm{d}n^p_{R_{p,i}}(t)}{\mathrm{d}t} = G^p_{R_{p,i-1}}(t) - G^p_{R_{p,i}}(t), \quad \left(\alpha^p_{R_{p,i}}\right) \quad \forall p \in \mathcal{P}, \ i \in [2, m(p)] \tag{13b}$$

$$\frac{\mathrm{d}E_w(t)}{\mathrm{d}t} = \sum_{p \in \mathcal{P}_w} q^p(t), \quad (\rho_w) \quad \forall w \in \mathcal{W} \tag{13c}$$

$$G^p_{R_{p,1}}\left(t + h_{R_{p,1}}\left(n_{R_{p,1}}(t)\right)\right)\left(1 + h'_{R_{p,1}}\left(n_{R_{p,1}}(t)\right) \cdot \dot{n}_{R_{p,1}}(t)\right) = q^p(t), \quad \left(\beta^p_{R_{p,1}}\right) \quad \forall p \in \mathcal{P} \tag{13d}$$

$$G^p_{R_{p,i}}\left(t + h_{R_{p,i}}\left(n_{R_{p,i}}(t)\right)\right)\left(1 + h'_{R_{p,i}}\left(n_{R_{p,i}}(t)\right) \cdot \dot{n}_{R_{p,i}}(t)\right) = G^p_{R_{p,i-1}}(t), \quad \left(\beta^p_{R_{p,i}}\right) \quad \forall p \in \mathcal{P}, \ i \in [2, \ m(p)] \tag{13e}$$

$$-q^p(t) \leq 0, \quad \left(\gamma^p_{R_{p,0}}\right) \quad \forall p \in \mathcal{P} \tag{13f}$$

$$-G^p_{R_{p,i}}(t) \leq 0, \quad \left(\gamma^p_{R_{p,i}}\right) \quad \forall p \in \mathcal{P}, \ i \in [1, \ m(p)] \tag{13g}$$

$$-n^p_{R_{p,i}}(t) \leq 0, \quad \left(\lambda^p_{R_{p,i}}\right) \quad \forall p \in \mathcal{P}, \ i \in [1, \ m(p)] \tag{13h}$$

$$\sum_{p \in \mathcal{P}} \left(q^p(t)\delta_{R_{p,1}, R_j} + \sum_{i=2}^{m(p)} G^p_{R_{p,i-1}}(t)\delta_{R_{p,i}, R_j} \right) \leq G^{max}_{R_j}, \quad \left(\zeta_{R_j}\right) \tag{13i}$$

$$E_w(T) = Q_w, \quad (\phi_w) \quad \forall w \in \mathcal{W} \tag{13j}$$

$$E_w(0) = 0, \quad \forall w \in \mathcal{W} \quad q^p(0) = 0, \quad \forall p \in \mathcal{P} \tag{13k}$$

where

$$n_{R_j}(t) = \sum_{p\in\mathcal{P}}\sum_{i=1}^{m(p)} n^p_{R_{p,i}}(t)\,\delta_{R_{p,i},R_j},\quad G_{R_j}(t) = \sum_{p\in\mathcal{P}}\sum_{i=1}^{m(p)} G^p_{R_{p,i}}(t)\,\delta_{R_{p,i},R_j},\quad Q_w = \sum_{p\in\mathcal{P}_w}\int_0^T q^p(t)\,\mathrm{d}t, \forall w\in\mathcal{W}$$

$$\delta_{R_{p,i},R_j} = \begin{cases} 1, & \text{if } R_{p,i} = R_j \\ 0, & \text{otherwise} \end{cases},\ \forall R_i\in\mathcal{R},\ \forall R_{p_i}\in\mathcal{R}_p,\ p\in\mathcal{P}$$

$n_{R_j}(t)$ and $n_{R_{p,i}}(t)$ denote the total accumulation state of region R_j in $\mathcal{R}$ and the total accumulation state of the i^{th} region along path p at time t, respectively; "$'$" denotes differentiation with respect to the associated function argument [1], while "$\cdot$" denotes differentiation with respect to time t. $\mathcal{P}_w$ is the set of paths connecting OD pair w and $E_w(t)$ is an extended state for flow conservation of OD pair w. $\delta_{R_{p,i},R_j}$ is the Kronecker Delta function. If the i^{th} region along path p, i.e., $R_{p,i}$, is the region $R_j\in\mathcal{R}$, then $\delta_{R_{p,i},R_j} = 1$; otherwise, $\delta_{R_{p,i},R_j} = 0$. The control variables are $\boldsymbol{q} = (q^p : p\in\mathcal{P})$, while $\boldsymbol{G} = (G^p_{R_{p,i}} : p\in\mathcal{P},\ i\in[1,m(p)])$, where $G^p_{R_{p,i}}$ denotes the outflow rate exiting $R_{p,i}$ from path p, $R_{p,i}$ is the i^{th} region along path p; $\boldsymbol{A} = (A_{R_j} : R_j\in\mathcal{R})$, where $A_{R_j} = \sum_{p\in\mathcal{P}}\left(q^p(t)\delta_{R_{p,1},R_j} + \sum_{i=2}^{m(p)} G^p_{R_{p,i-1}}(t)\delta_{R_{p,i},R_j}\right)$ denotes the total entry rate to region R_j. (13a)-(13b) are network traffic dynamics corresponding to (6a)-(6b). (13d)-(13e) are flow propagation constraints, which is an equivalent version of (4). Note for any given $q^p(t)\geq 0,\ \forall p\in\mathcal{P}$, we have $G^p_{R_i}(t)\geq 0,\ \forall p\in\mathcal{P},\ \forall R_i\in\mathcal{R}$ and $n_{R_j}(t)\geq 0,\ \forall R_j\in\mathcal{R}$ under FIFO condition. (13c) and (13j) are flow conservation constraints, while (13k) specifies the initial conditions. (13i) restricts that the inflow to a region should be less than its maximum service rate or capacity. Variables in brackets of (13a)-(13j) are Lagrange multipliers associated with the corresponding constraints. ∎

To derive the necessary conditions for DVI (12), a mathematical convenience for analyzing the necessary conditions is to restate it as an optimal control problem as:

$$\min J = \sum_{w\in\mathcal{W}}\sum_{\forall p\in\mathcal{P}_w}\int_0^T \Psi^p(t,\boldsymbol{n}^*)\,q^p(t)\,\mathrm{d}t \tag{14}$$

subject to the flow propagation, state and input constraints specified in (13a)-(13k).

Proposition 3.1. The necessary condition for **DUE I** with simultaneous route choice and departure time choice can be stated as:

$$q^p(t)\begin{cases} >0 \Rightarrow \Psi^p(t,\boldsymbol{n}^*) + l_p(t) = \phi_w, \\ =0 \Rightarrow \Psi^p(t,\boldsymbol{n}^*) + l_p(t) > \phi_w, \end{cases}\quad \forall p\in\mathcal{P}_w\ \ w\in\mathcal{W}$$

where $\boldsymbol{n}^*$ is the optimal state of the **DUE I**; ϕ_w specifies the travel cost under the DUE condition determined by the fixed total travel demand Q_w.

$$l_p(t) = \sum_{i=1}^{m(p)}\sum_{R_j\in\mathcal{R}}\zeta_{R_j}\left(\tau^p_{R_{p,i-1}}(t)\right)\delta_{R_{p,i},R_j} \triangleq \sum_{i=1}^{m(p)}\zeta_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}(t)\right)$$

represents the dynamic external cost imposed on travelers for their presence on the regions along the trip trajectory, wherein $\zeta_{R_i}(\cdot)$ is the Lagrange multiplier associated with the upper-bound of the inflow to region

[1] It is assumed that all network delay functions are differentiable with regard to their own arguments.

R_i due to its limited network capacity. Note that there is one and only one region $R_j \in \mathcal{R}$ would be the i^{th} region along path p, i.e., $R_{p,i}$, $\sum_{R_j \in \mathcal{R}} \delta_{R_{p,i},R_j} = 1$. To save notations, we have defined $\zeta_{R_{p,i}}(\cdot) \triangleq \sum_{R_j \in \mathcal{R}} \zeta_{R_j}(\cdot)\delta_{R_{p,i},R_j}$ to evaluate the additional cost caused by the regional inflow restrictions along path p. $\tau^p_{R_{p,i}}(t)$ denotes the exit time from the i^{th} region $R_{p,i}$ and also the entry time to the $i+1^{th}$ region $R_{p,i+1}$ of vehicles departing from the origin at time t using path p, defined by (8). ∎

Proof of Proposition 3.1. See Appendix C. ■

Remark 3.1. The inflow capacity constraint has been widely adopted in the DTA literature by tracing the flow propagation in the network loading numerically, see, e.g., Yildirimoglu et al. (2015). To the best of our knowledge, (13i) is the first analytical formulation of such inflow capacity constraint for general networks thanks to the definition of $\delta_{R_{p,i},R_j}$. In this sense, **DUE I** can be regarded as the first analytical equilibrium condition for DUE problems with inflow capacity constraints. ∎

Note that the network output function of the MFD system with time delay, i.e., the trip completion rate $G_{R_i}(n_{R_i})$ of region R_i, endogenously depends on the network state as a concave function of network accumulation, with $G_{R_i}(n_{R_i}) = 0$ when $n_{R_i} = n^{jam}_{R_i}$. The phenomenon that accumulation exceeding sustainable point leads to zero exit function is recognized as "gridlock" in the MFD literature (Daganzo, 2007; Mahmassani et al., 2013), which is not desired for the sake of urban mobility. One possible alternative to avoid the "gridlock" is to impose $n_{R_i(t)} \leq n^{jam}_{R_i} - \epsilon_{R_i}$ such that the MFD is well-defined from theoretical consideration and the network is protected from gridlock (Zhong et al., 2018b). On the other hand, the control objective is usually to prevent the network from over-saturation. Therefore, it is reasonable to further incorporate the following state constraint in the **DUE I:**

Problem 3.2. DUE II: (12) s.t. (13a)-(13k) and

$$n_{R_j}(t) \leq n^{cr}_{R_j}, \quad \left(\eta_{R_j}\right) \quad \forall R_j \in \mathcal{R} \tag{15}$$

(15) is saturation constraint imposed by MFD determined by physical characteristics of the network, e.g., the critical point to protect the region from over-saturation. Imposing the accumulation constraint (15) can ensure the FIFO condition using nonlinear travel time functions as outlined in Appendix E. ∎

Corollary 3.1. The necessary condition for **DUE II** with simultaneous route choice and departure time choice can be stated as follows:

$$q^p(t) \begin{cases} > 0 \Rightarrow \Psi^p(t, \boldsymbol{n}^*) + l_p(t) = \phi_w, \\ = 0 \Rightarrow \Psi^p(t, \boldsymbol{n}^*) + l_p(t) > \phi_w, \end{cases} \quad \forall p \in \mathcal{P}_w \quad w \in \mathcal{W}$$

where $\boldsymbol{n}^*$ is the optimal state of the **DUE II** for OD pair w; ϕ_w specifies the travel cost under the DUE condition determined by the fixed total travel demand Q_w.

$$\begin{aligned} l_p(t) &= \sum_{i=1}^{m(p)} \int_{\tau^p_{R_{p,i-1}}}^{\tau^p_{R_{p,i}}} \left(\sum_{R_j \in \mathcal{R}} \eta_{R_j}(u)\delta_{R_{p,i},R_j} \right) \mathrm{d}u + \sum_{i=1}^{m(p)} \sum_{R_j \in \mathcal{R}} \zeta_{R_j}\left(\tau^p_{R_{p,i-1}}(t)\right)\delta_{R_{p,i},R_j} \\ &= \sum_{i=1}^{m(p)} \int_{\tau^p_{R_{p,i-1}}}^{\tau^p_{R_{p,i}}} \eta_{R_{p,i}}(u)\,\mathrm{d}u + \sum_{i=1}^{m(p)} \zeta_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}(t)\right) \end{aligned}$$

represents the dynamic external cost imposed on travelers for their presence on the regions along the trip trajectory, wherein $\eta_{R_i}(t)$ denotes the Lagrange multiplier associated with the state constraint imposed on the accumulation of region R_i, and $\zeta_{R_i}(\cdot)$ is defined in Proposition 3.1. Similar to the definition of $\zeta_{R_{p,i}}(\cdot)$ in Proposition 3.1, we have defined $\eta_{R_{p,i}}(\cdot) \triangleq \sum_{R_j \in \mathcal{R}} \eta_{R_j}(\cdot)\delta_{R_{p,i},R_j}$ to save notations. $\int_{\tau^p_{R_{p,i-1}}}^{\tau^p_{R_{p,i}}} \eta_{R_{p,i}}(u)\,\mathrm{d}u$ is the penalty cost to access the i^{th} protected region $R_{p,i}$ along path p. $\tau^p_{R_{p,i}}(t)$ denotes the exit time from the i^{th} region $R_{p,i}$ and also the entry time to the $i+1^{th}$ region $R_{p,i+1}$ of vehicles departing from the origin at time t using path p, defined by (8). ∎

Proof of Corollary 3.1. See Appendix D. ■

Remark 3.2. Following Carey et al. (2014), we show in Appendix E that the nonlinear travel time function $h(n(t))$ fulfils the FIFO condition if $n \leq n_{cr}$ is enforced. As a matter of fact, it is commonly adopted in the MFD literature that the desired equilibrium state is set to the critical accumulation n_{cr} (or close to n_{cr} from the left-hand side). Therefore, we impose this constraint on the network state in the **DUE II**.

In the literature, perimeter control, that gates inter-boundary transferring traffic flows, would result in virtual queues that do not interact with the rest of the traffic at regional boundaries. Such boundary queues create unmodeled dynamics that may affect the existence of MFDs and would further induce negative effects such as spillback in reality (Kouvelas et al., 2017). To store the virtual queue, Zhong et al. (2018b) introduced the concept of "source" with limited capacity such that the demand cannot be received by the network due to congestion is buffered in the source. The dynamics of the source is similar to the deterministic queueing (or bottleneck) model and thus the vertical queue. The interface between the source and the MFD system is described by a set of demand and supply reaction law. From both theoretical and practical points of view, physical queue and thus more complex dynamics are necessary to describe the boundary queueing behavior while how to determine the bounds of physical queue, especially the feasible upper bound, remains unsolved (Haddad, 2017b). On the other hand, under such circumstances, a large parking hub would be required to accommodate the queueing vehicles which is unlikely to be attainable in dense urban areas.

In this paper, we propose the schedule delay cost approach to incorporate the cost induced by the over-saturated traffic without introducing additional complex traffic dynamics. The schedule delay cost consists of the induced penalty cost by the constraints and that of the preferable arrival. This schedule delay cost helps resolving the missing region-to-region boundary queue problem in the sense that the schedule delay cost would force travelers to change their route choice and departure time choice to avoid the additional cost to protect the network from over-saturation. From a practical point of view, if the path flows of one assignment would cause some regions being over-saturated that the inflows to the region cannot be received, the path flows of this assignment cannot be the equilibrium assignment since the flow propagation is interrupted by congestion. Such an equilibrium seeking process will continue until such flow rejection phenomenon will not occur, i.e., the flow constraints are well satisfied. Otherwise, if these constraints are activated, i.e., the equal sign in (13i) and (15) hold, the associated Lagrange multipliers η and ζ would introduce large additional travel cost to travelers who access the network during saturated period, and thus forcing them to change their routes and departure times, which in turn prohibit the region-to-region boundary queue from occurring. In other words, the inflow capacity and accumulation constraints give rise to an alternative approach to capturing the hypercongestion represented by the downward sloping part of the MFD while maintaining the FIFO using nonlinear travel time functions. ∎

Remark 3.3. In the context of road pricing, link capacity constraint is usually imposed while the Lagrange multipliers associated with the capacity constraints are regarded as the "shadow prices" (Yang and Huang, 2005). Link tolls thus obtained are implemented to prevent the links from over-saturation. In particular, the designed criterion of dynamic congestion charging is to eliminate queuing via the optimal time-varying (link) tolls. The additional cost caused by the inflow capacity and network accumulation constraints in this paper can be regarded as a network-level extension of the above shadow price mechanism. ∎

4. Solution algorithms

This section presents numerical treatings regarding the proposed DUE problem. As discussed in Zhong et al. (2011); Zhong (2011), the DUE problem cannot be solved by the optimal control formulation directly but through an iterative optimal control problem. On the other hand, to adopt off-the-shelf nonlinear optimization algorithms to solve the related optimal control problems, it is necessary to apply time-discretization to yield a finite dimensional nonlinear optimization approximation of the original infinite dimensional optimal control problems.

4.1. Optimal control reformulations of the DUE

Introducing a vector representation of the path flows $\boldsymbol{v} \geq 0$ and based on the DVI formulation (12), the **DUE I** can be solved by $\boldsymbol{q} = \arg\min_{\boldsymbol{v}} \left\{ \frac{1}{2} \left\| \boldsymbol{q} - \varpi \Psi(t,\, \boldsymbol{n}(\boldsymbol{q})) - \boldsymbol{v} \right\|^2 : \boldsymbol{v} \in \bar{\Lambda} \right\}$

$$\min_{\boldsymbol{v} \in \bar{\Lambda}} J(\boldsymbol{v}) = \frac{1}{2} \int_0^T \left[\boldsymbol{q} - \varpi \Psi(t,\, \boldsymbol{n}(\boldsymbol{q})) - \boldsymbol{v} \right]^T \left[\boldsymbol{q} - \varpi \Psi(t,\, \boldsymbol{n}(\boldsymbol{q})) - \boldsymbol{v} \right] \mathrm{d}t, \quad \forall \boldsymbol{q} \in \bar{\Lambda} \tag{16}$$

where

$$\bar{\Lambda} = \Lambda \bigcap \tilde{\Lambda}, \quad \Lambda = \left\{ \boldsymbol{q} : \boldsymbol{q} \geq 0, \ \sum_{p \in \mathcal{P}_w} \int_0^T q^p(t)\, \mathrm{d}t = Q_w, \ \forall p \in \mathcal{P}_w, \ \forall w \in \mathcal{W}, \ \forall t \in [0, T] \right\}$$

$$\tilde{\Lambda} = \{ \boldsymbol{q} : \boldsymbol{q} \geq 0, \ \boldsymbol{A} \in \Sigma_A, \ \boldsymbol{n} \in \Sigma_n \}, \quad \Sigma_A \triangleq \{ \boldsymbol{A}(t) : 0 \leq \boldsymbol{A}(t) \leq \boldsymbol{G}^{max} \}, \quad \Sigma_n \triangleq \{ \boldsymbol{n}(t) : 0 \leq \boldsymbol{n}(t) \}$$

while for **DUE II**

$$\Sigma_n \triangleq \{ \boldsymbol{n}(t) : 0 \leq \boldsymbol{n}(t) \leq \boldsymbol{n}^{cr} \}$$

$\tilde{\Lambda}$, Σ_A and Σ_n are compact.

Similar to Zhong et al. (2011), this infinite dimensional mathematical programming can be solved as an iterative optimal control problem (or a fixed point algorithm) as depicted in Algorithm 1, where $\Psi(q)$ is replaced by the effective path delay operators $\Psi^p(t,\, n)$, $\forall p \in \mathcal{P}$, see e.g., (16); $\dot{n} = \tilde{f}(\cdot)$ denotes the network traffic dynamics with the flow propagation constraints (13a)-(13e). Since the network accumulations are state variables in this optimal control problem, $\dot{n}$ can be explicitly written. With travel time function defined by (8), the flow propagation constraints (13d)-(13e) are substituted into (13a)-(13b) to update the state variables. $\mathcal{M}_1(\boldsymbol{q}, t) \leq 0$ is a compact representation of the set

$$\left\{ \boldsymbol{q} : -\boldsymbol{q}(t) \leq 0, \quad \sum_{p \in \mathcal{P}_w} \int_0^T q^p(t)\, \mathrm{d}t = Q_w, \ \forall p \in \mathcal{P}_w, \ \forall w \in \mathcal{W}, \ \forall t \in [0, T] \right\}.$$

Algorithm 1 Fixed point algorithm for the DUE problem

Input: Maximum iteration number I_{max}, tolerance $\varepsilon \in R_+$

Output: Inflow profiles under DUE $\boldsymbol{q}^*$

1: Initialization: set $l = 1$ and identify an initial feasible solution $\boldsymbol{q}^l \in \bar{\Lambda}$, $\varpi \in R_+$

2: **while** $l \leq I_{max}$ and $\frac{\|\boldsymbol{q}^{l+1} - \boldsymbol{q}^l\|}{\|\boldsymbol{q}^l\|} > \varepsilon$ **do**

3: Optimal control subproblem. Solve

$$\min_{\boldsymbol{v}} J^l(\boldsymbol{v}) = \frac{1}{2}\int_0^T \left[\boldsymbol{q}^l - \varpi\Psi\left(\boldsymbol{n}(\boldsymbol{q}^l)\right) - \boldsymbol{v}\right]^T \left[\boldsymbol{q}^l - \varpi\Psi\left(\boldsymbol{n}(\boldsymbol{q}^l)\right) - \boldsymbol{v}\right] \mathrm{d}t \triangleq \int_0^T F_2\left(\boldsymbol{n}^l, \boldsymbol{q}^l, \boldsymbol{v}\right) \mathrm{d}t$$

$$\text{s.t.} \quad \dot{\boldsymbol{n}} = \tilde{f}(\boldsymbol{n}, \boldsymbol{v}, t),\ \mathcal{M}_1(\boldsymbol{v}, t) = 0,\ \mathcal{M}_2(\boldsymbol{n}, t) \leq 0,\ \mathcal{M}_3(\boldsymbol{v}, t) \leq 0,\ -\boldsymbol{v} \leq 0,\ \forall \boldsymbol{q} \in \bar{\Lambda},\ \forall t \in [0, T],\ \boldsymbol{n}(0) = 0$$

4: $l \leftarrow l + 1$, $\boldsymbol{q}^l = \boldsymbol{v}^*$

5: **return** $\boldsymbol{q}^* \approx \boldsymbol{q}^l$

The inequality constraint $\mathcal{M}_2(\cdot)$ denotes the pure state constraints (13h) and (15). $\mathcal{M}_3(\boldsymbol{q}, t)$ denotes the regional inflow constraints per (13i). Although Friesz and Mookherjee (2006) analyzed the convergence of such a fixed point algorithm for an abstract feasible region with certain properties, the convergence is still generally heuristic since convergence conditions are unlikely to be verified for general traffic networks, e.g., strong monotone of $F_2\left(\boldsymbol{n}^l, \boldsymbol{q}^l, \boldsymbol{v}\right)$.

4.2. Time-discretization of the MFD with time delay dynamics

To solve the optimal control subproblems with state and/or control constraints and state-dependent time delays as indicated by Algorithm 1 via nonlinear programming algorithms, e.g., sequential quadratic programming (SQP), it is necessary to discretize the dynamics of MFD with time delay first. A simple method is to divide the time horizon into K uniform intervals with $K = \lfloor T/\Delta t \rfloor$, where floor function $\lfloor x \rfloor$ gives the greatest integer that is less than or equal to x and Δt is the time increment. This idea leads to the following numerical network loading algorithm which essentially consists of flow propagation (Algorithm 2 and Algorithm 3) and path travel time calculation.

4.3. Solution algorithm for the DUE problem and the numerical scheme for dynamic externalities

After the network loading model is discretized in time, applying time discretization to the objective function and other constraints of the DUE problem yields the following NLP approximation of the optimal control problem as described in Algorithm 1:

$$\begin{aligned}
\min_{z} J^l(z) &= \sum_{i=0}^{K-1} F_2(\boldsymbol{n}_l, \boldsymbol{q}_l, \boldsymbol{v}_i)\Delta t \\
\text{s.t.} \quad \boldsymbol{n}_{i+1} &= \boldsymbol{n}_i + \Delta t \cdot f_0(\boldsymbol{n}_i, \boldsymbol{v}_i, t_i), \quad i = 0, 1, \ldots, K-2 \\
\boldsymbol{n}_i &= \boldsymbol{n}_i(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{i-1}, t_i), \quad i = 1, 2, \ldots, K-1, \quad \boldsymbol{n}_0 = 0 \\
\mathcal{M}_1(\boldsymbol{v}_i, t_i) &= 0,\ \mathcal{M}_2(\boldsymbol{n}_i, t_i) \leq 0,\ \mathcal{M}_3(\boldsymbol{v}_i, t_i) \leq 0,\ -\boldsymbol{v}_i \leq 0,\ i = 0, 1, \ldots, K-1 \\
\text{where} \quad z &= [v_0, v_1, \ldots v_{K-1}]^T
\end{aligned}$$

which is solved by constrained nonlinear programming, e.g., SQP, in the inner loop of Figure 3. While in the outer loop, an optimum is achieved if $\frac{\|\boldsymbol{q}^{l+1} - \boldsymbol{q}^l\|}{\|\boldsymbol{q}^l\|} \leq \varepsilon$, where ϵ is preset positive tolerance. After the NLP

Algorithm 2 Network loading

1: Initialization: for all paths, initialize the entry time $\tau^p_{R_{p,0}}(k_t)$ $(k_t = 1, 2, ..., K, K+1)$ and departure rate $q^p_{k_t}$ $(k_t = 1, 2, ..., K)$ for all time interval index, by

$$\tau^p_{R_{p,0}}(k_t) = (k_t - 1)\Delta t,\ q^p_{k_t} = q^p\left(\tau^p_{R_{p,0}}(k_t)\right)$$

initialize the cumulative inflow of the first region of path p, $Q^p_{R_{p,0}}(k_t)$ by

$$Q^p_{R_{p,0}}(k_t) = \sum_{k=1}^{k_t-1} q^p_k \Delta t$$

where k_t is the time interval index

2: Load each region in turn, until all region have been loaded

2.1 Initialization: set the current region $R_j = R_1$, the number of loaded regions $m = 0$
2.2 If region R_j have been loaded, go to step 2.4
2.3 If the entry time and the cumulative inflow of all path flows into current region R_j have been loaded, load region R_j by Algorithm 3; then we obtain the exit time $\tau^p_{R_j}(k_t)$ and the cumulative outflow $Q^p_{R_j}(k_t)$ of all path flows from R as the entry time and cumulative inflow to the downstream region along the path. The number of loaded regions $m = m + 1$;
2.4 If $R_j = R_M$, the current region is the last region, reset $R = R_1$, otherwise $R_j = R_{j+1}$
2.5 If $m = M$, all region have been loaded; otherwise, go to step 2.2

Algorithm 3 Region loading

1: Initialization: R_j: the region to be loaded; $\hat{\mathcal{P}} = \{p \subseteq \mathcal{P} | R_j \in \mathcal{R}_p\}$: the set of all paths passed by the region R_j, and resort all paths in $\hat{\mathcal{P}}$ with new label $\hat{p}_1, \hat{p}_2, ..., \hat{p}_r$, r is the number of elements in $\hat{\mathcal{P}}$

2: Initialization: the number of loaded paths $s = 0$

3: Load the region at once according to the entry time of all paths in $\hat{\mathcal{P}}$

3.1 Initialization: Entry time $\tau^{\hat{p}_i}_{in}$, cumulative entry flow $Q^{\hat{p}_i}_{in}$ and time label $k^{\hat{p}_i}_t = 1$ for $i = 1, 2, ..., r$.
3.2 Find the earliest time corresponding to all the time labels $\hat{p}^* = \arg_i \min(\tau^{\hat{p}_i}_{in}(k^{\hat{p}_i}_t))$
3.3 Calculate the accumulation state of a region $\mathcal{N}$ at time $\tau^{\hat{p}^*}_{in}(k^{\hat{p}^*}_t)$ by Algorithm 4
3.4 Calculate region travel time $h_{R_j}(\mathcal{N})$, and update the exit time $\tau^{\hat{p}^*}_{out}(k^{\hat{p}^*}_t) = \tau^{\hat{p}^*}_{in}(k^{\hat{p}^*}_t) + h_{R_j}(\mathcal{N})$, and cumulative exit flow $Q^{\hat{p}_i}_{out}(k^{\hat{p}^*}_t) = Q^{\hat{p}_i}_{in}(k^{\hat{p}^*}_t)$
3.5 If $k^{\hat{p}^*}_t$ is the last time point, $s = s + 1$
3.6 If $s < r$, go to Step 3.2

4: Update the exit time and cumulative exit flow to the next region along path for all paths, and stop

Algorithm 4 Calculate accumulation state of a region

1: Initialization: R_j, $\hat{\mathcal{P}}$, r, entry time $\tau_{in}^{\hat{p}_i}$, cumulative entry flow $Q_{in}^{\hat{p}_i}$, and exit time $\tau_{out}^{\hat{p}^*}$, cumulative exit flow $Q_{out}^{\hat{p}_i}$

2: Initialize $t = \tau_{in}^{\hat{p}^*}(k_t^{\hat{p}^*})$

3: Calculate the cumulative inflow Q_{in} and cumulative outflow Q_{out} at time t for all paths in $\hat{\mathcal{P}}$

3.1 Initialization: path index $i = 1$

3.2 Calculate cumulative inflow of path $\hat{p}_i$
if $t < \tau_{in}^{\hat{p}_i}(1)$, no vehicle enters the region along path $\hat{p}_i$. $Q_{in} = Q_{in} + 0$;
if $t > \tau_{in}^{\hat{p}_i}(K)$, all flows along path $\hat{p}_i$ have entered the region. $Q_{in} = Q_{in} + Q_{in}^{\hat{p}_i}(K)$;
if $\tau_{in}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i}) < t \leq \tau_{in}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i} + 1)$, calculate the cumulative inflow by interpolation

$$Q_{in} = Q_{in} + Q_{in}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i}) + \frac{Q_{in}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i} + 1) - Q_{in}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i})}{\tau_{in}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i} + 1) - \tau_{in}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i})}(t - \tau_{in}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i}))$$

3.3 Calculate cumulative outflow of path $\hat{p}_i$
if $t < \tau_{out}^{\hat{p}_i}(1)$, no vehicle exits the region along path $\hat{p}_i$. $Q_{out} = Q_{out} + 0$;
if $t > \tau_{out}^{\hat{p}_i}(K)$, all flows along path $\hat{p}_i$ have exited the region. $Q_{out} = Q_{out} + Q_{out}^{\hat{p}_i}(K)$;
if $\tau_{out}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i}) < t \leq \tau_{out}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i} + 1)$, calculate the cumulative outflow by interpolation

$$Q_{out} = Q_{out} + Q_{out}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i}) + \frac{Q_{out}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i} + 1) - Q_{out}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i})}{\tau_{out}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i} + 1) - \tau_{out}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i})}(t - \tau_{out}^{\hat{p}_i}(\tilde{k}_t^{\hat{p}_i}))$$

3.4 If $i < r$, $i = i + 1$, go to step 3.2

4: Accumulation state of a region $\mathcal{N} = Q_{in} - Q_{out}$

Algorithm 5 Calculate externality l_p

1: Initialization: p, and set entry time index $k_t = 1$, $i = 0$

2: Initialize $l_p(k_t) = 0$

3: Calculate additional cost for the i^{th} region along path p

$$l_p(k_t) = l_p(k_t) + \int_{\tau_{R_{p,i-1}}^p(k_t)}^{\tau_{R_{p,i}}^p(k_t)} \eta_{R_{p,i}}(u)\,\mathrm{d}u + \zeta_{R_{p,i}}\left(\tau_{R_{p,i-1}}^p(k_t)\right)$$

4: If $i < m(p)$, $i = i + 1$, go to Step 3

5: If $k_t < K$, $k_t = k_t + 1$, go to Step 2; otherwise stop.

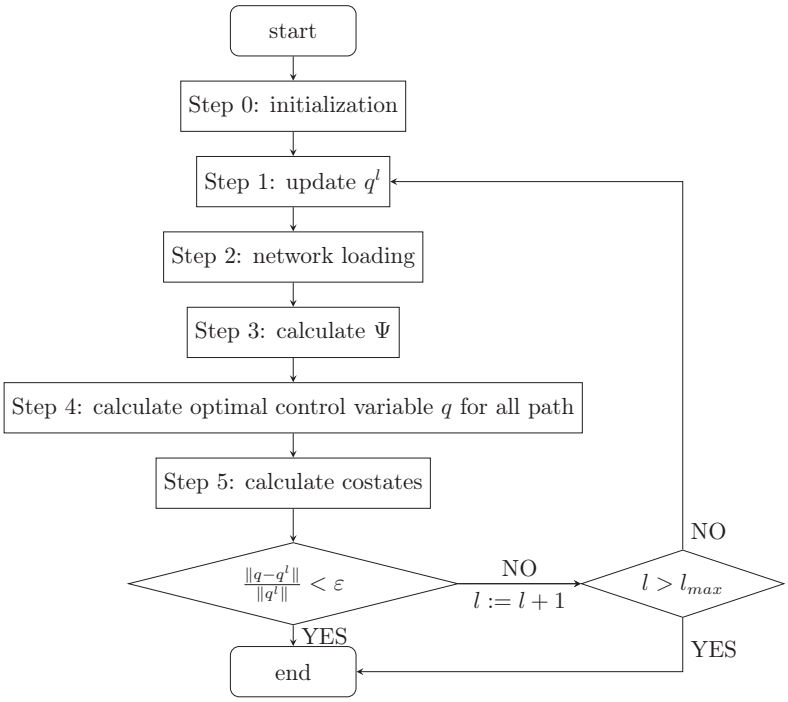


Figure 3: Flowchart of the solution algorithm

is solved, we need to evaluate the dynamic external costs to obtain the DUE equilibrium condition [2].

Based on above time discretization, we can evaluate the dynamic external costs (by Algorithm 5) [3] in the DUE equilibrium condition.

5. Numerical examples

For illustrative purpose, we consider three scenarios with respect to the OD pairs, i.e., within region trips whose origins and destinations are in the same region in Subsection 5.1, the Braess' network case whose origins and destinations are in different regions in Subsection 5.2, and two OD pairs sharing overlapping paths in Subsection 5.3, respectively.

5.1. Within-region trips

Consider an urban network admitting a well-defined MFD widely used in the literature (Geroliminis et al., 2013), e.g., $G(n) = (1.4877 \times 10^{-7} \times n^3 - 2.9815 \times 10^{-3} \times n^2 + 15.0912 \times n)/3600$, with $G^{max} = 6.3$ (unit/time). The nonlinear network travel time function is described by $h(t) = n(t)/G(n(t))$, while the network free flow travel time is $h(0) = 238$ (times). The total amount of within-region demand to be assigned is $Q_{od} = 1500$ (units). The planning horizon is set at $T = 800$ (times) to ensure all travel demand

[2] Solving DUE I and II obtains optimal control variable $\boldsymbol{q}^*$ and state variable $\boldsymbol{n}^*$. Note the Lagarange multipliers in externality evaluation correspond to problem (14), optimal $\boldsymbol{n}^*$ was substituted into (14) to extract Lagarange multipliers η and ζ.

[3] Algorithm 5 can be used to evaluate the dynamic eternal costs of DUE I and II wherein $\eta = 0$ in DUE I.

can be served. The upper bound of the inflow rate is set as the maximum capacity at 6.3 (unit/time). Early/late arrival penalty is defined as

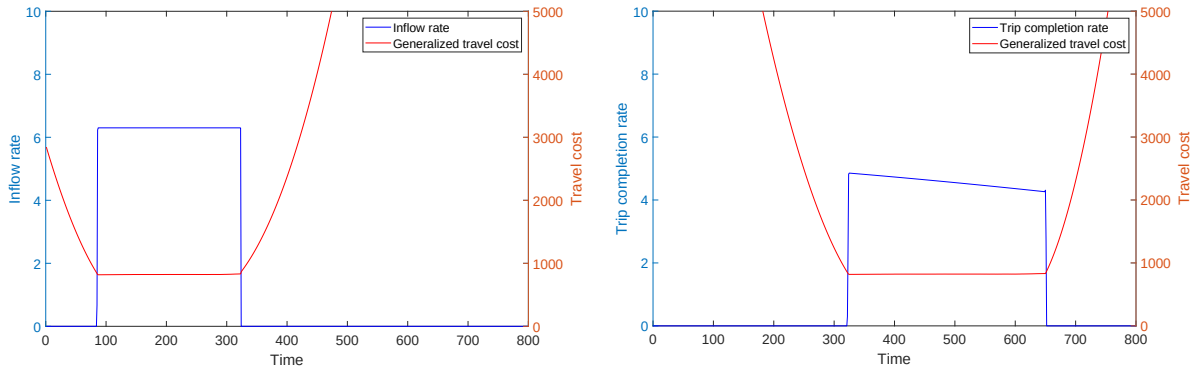
$$\kappa[\chi] = \begin{cases} 0.1\,(t + h(n) - t_{de})^2\,, & t + h(n) < t_{de} \\ 0, & t_{de} \le t + h(n) \le t_{dl} \\ 0.2\,(t + h(n) - t_{dl})^2\,, & t + h(n) > t_{dl} \end{cases} \tag{17}$$

with expected arrival time ranging from $[t_{de},\ t_{dl}] = [400,\ 600]$ (times).

For the within-region trips, travelers do not have route choice but departure time choice. Solving the DUE, the path inflow rate against the corresponding travel cost over time is depicted in Figure 4(a). As shown in Figure 4(a), the generalized travel cost (remains constant) and the path flow during the departure window well satisfy the DUE condition. The earliest departure of the DUE case is at around 85 (times) and lasts until 323 (times). Figure 4(b) depicts the path trip completion rate against travel cost. The arrival window is larger than the preferred arrival window. This is because travelers choose their departure times to minimize the generalized travel cost (caused by saturated traffic condition and traffic control). Figure 4(c) and Figure 4(d) respectively illustrate the network accumulation and the cumulative completed trips indicating the flow conservation.

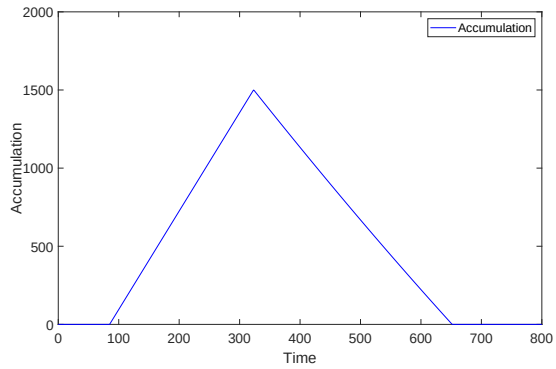
Note that there is no route choice but departure time choice for the within-region case, it is similar to the perimeter control located at a regional border, that manipulates the transfer flows across the border to optimize the regional operational performance, with control input saturation as investigated in Haddad (2017a,b). The inflow profile depicted in Figure 4(a) is similar to the Bang-bang like optimal control law obtained in the conventional optimal control formulations of gating control (Haddad, 2017a; Zhong et al., 2018a). Despite their similarity, we would like to point out that there is no guarantee that the optimal departure rate will be the Bang-bang law as achieved in the conventional optimal control formulations of gating control due to the following two reasons. First, the objective of the DUE problem is that travelers choose their route and departure time to minimize their travel cost while the optimal perimeter control is to maximize network throughput. Moreover, the MFD dynamics adopted in the conventional optimal control formulations does not involve the time-varying state delay.

As highlighted in Remark 3.2 that unlike the conventional approaches that ignore the boundary queues or require additional dynamics to describe the boundary queues, the proposed model introduces additional cost caused by schedule delay and penalty cost (induced by state and input constraints) to prevent both the boundary queues and over-saturated traffic conditions. Travelers choose their departure time to avoid the schedule delay penalty thus the trip completion profiles spread along the expected arrival time as shown in Figure 4(b). In the meanwhile, travelers manage to avoid the additional cost induced by the network capacity, i.e., the upper bound on the inflow profile, to minimize their generalized travel cost as indicated in Figure 4(a). This can avoid the difficulties in determining the steady state and the maximum admissible values of the boundary queues, which was highlighted as a future research topic in Haddad (2017b). One implication of the dynamic additional cost is that one can further devise demand management schemes such as road pricing to charge the travelers for accessing the over-saturated regions. Assume the VOT is 100 HKD/hr for all travelers, a dynamic tolling scheme can be obtained by converting the additional cost into monetary cost as shown in Figure 4(e). The toll is implemented during travelers' departure window to force them to change their departure time choice and thus protecting the network from saturation. The toll level is reasonable. The additional cost can be regarded as a network-level extension of the shadow price mechanism as discussed in Remark 3.3.

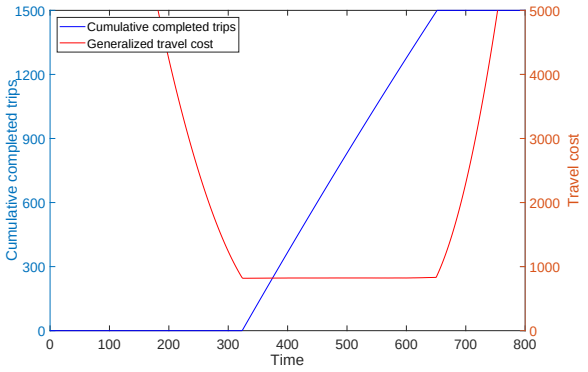


(a) Inflow rate and generalized travel cost

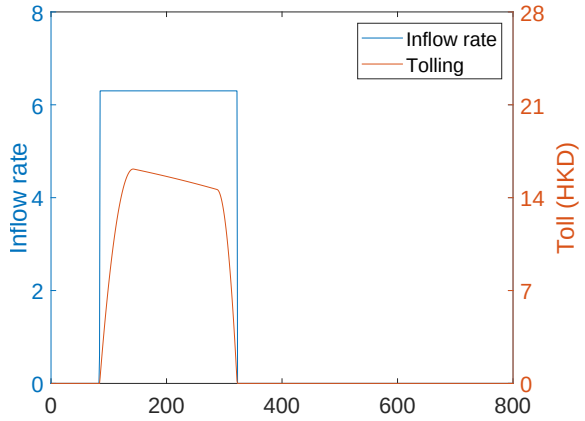
(b) Trip completion rate and generalized travel cost



(c) Accumulation state



(d) Cumulative completed trips



(e) Level of charge

Figure 4: An example of within-region trips with nonlinear travel time function

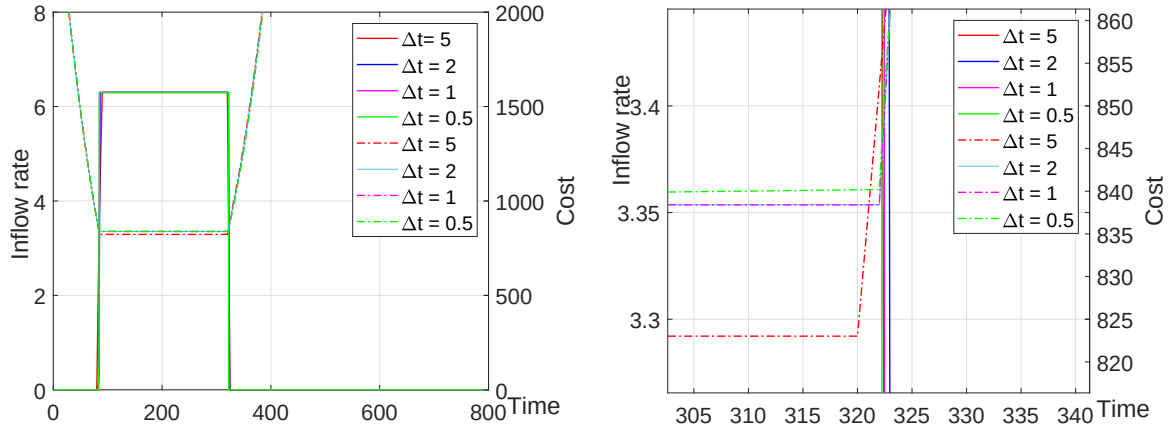


Figure 5: Effects of time-discretization: solid lines and dot dashed lines indicate inflow rate and generalized travel cost with time discretization equal to 5, 2, 1, 0.5, respectively.

5.1.1. The effects of time-discretization

The granularity of time discretization not only affects the computational efficiency of the algorithm but also the convergence and accuracy. To see this, the once-at-a-time sensitivity analysis is employed. Figure 5 depicts the path inflow against the generalized travel cost with respect to time increments of discretization $\Delta t = 0.5, 1, 2, 5$, respectively. As implied by Proposition 3.1, there shall be no travelers choosing to depart when the generalized travel exceeds the equilibrium cost. This equilibrium condition is well satisfied for these time increments of discretization as indicated in Figure 5(a). On the other hand, there is a time interval (smaller than the time increment $\Delta t = 5$) wherein the departure rate is positive even though the generalized cost exceeds the equilibrium cost as shown in Figure 5(b). This is due to the resolution of time-discretization and thus can be regarded as a tolerable error of the departure time window. Similarly, such a resolution of time-discretization could also contribute an error around $\Delta t = 5$ unit-times in the equilibrium cost. Therefore, we can conclude that the equilibrium condition is better satisfied as the granularity gets smaller. However, smaller time-discretization could lead to better approximation of the theoretical DUE condition at the price of increasing the computational burden.

5.2. A Braess like network

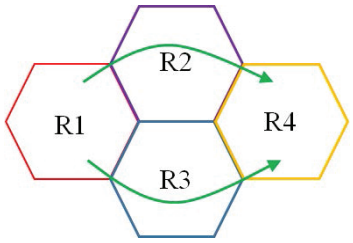


Figure 6: An example of a Braess like network

Consider a Braess like network consisting of 4 regions with a single OD pair tied by 2 paths as shown in Figure 6. The OD pair from R_1 to R_4 is connected by 2 paths, $R_1 \rightarrow R_2 \rightarrow R_4$ (Path 1) and $R_1 \rightarrow R_3 \rightarrow R_4$

(Path 2) while the configuration of each region with nonlinear network delay function is specified in Table 1, with $n \in [0, 4000]$. The total travel amount between R_1 and R_4 is $Q_{od} = 1500$ (units). The planning horizon is $T = 800$ (unit-times) with regional inflow rate restricted to be less than or equal to 12.6 (unit/time). Expected arrival time interval is specified as $[t_{de}, t_{dl}] = [500, 700]$ (times).

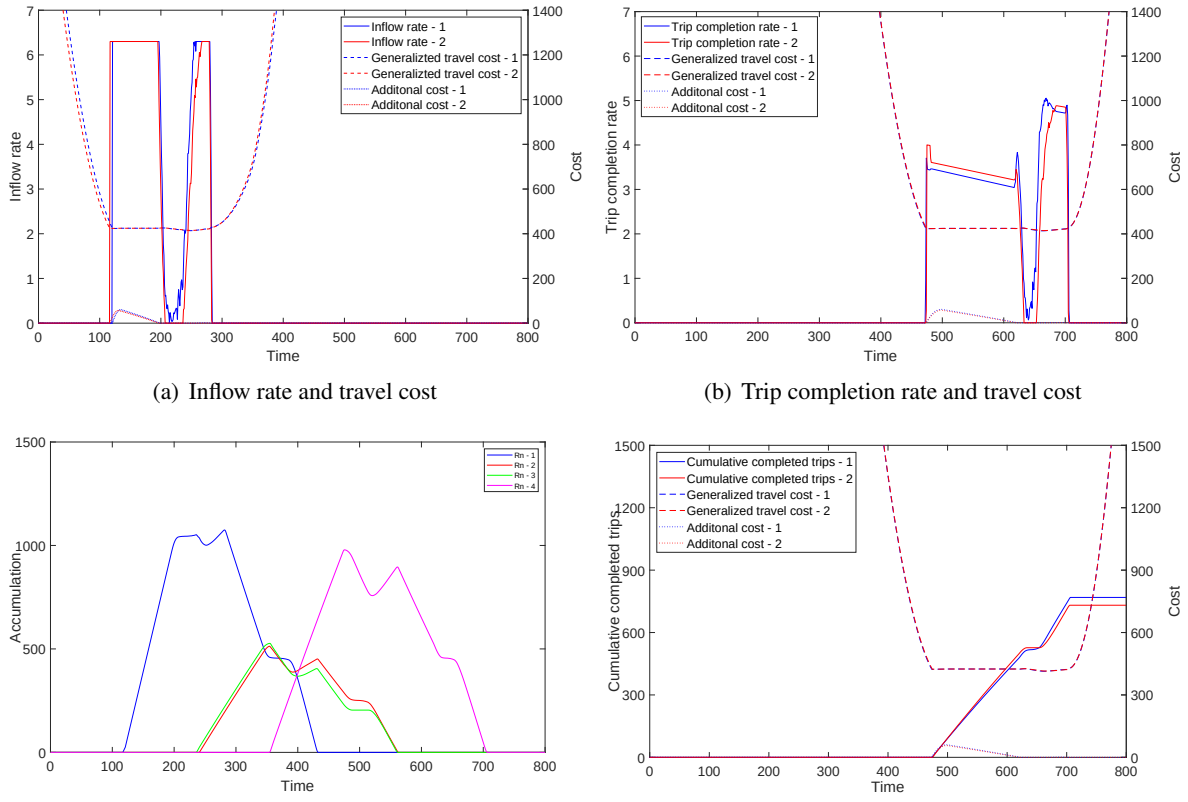
Table 1: Network configurations: Nonlinear network travel time function

	Network configuration	Nonlinear network delay $h_i(n_i(t))$
R_1	$a_1 = 2 \times 1.4877 \times 10^{-7}$, $b_1 = 2 \times 2.9815 \times 10^{-3}$, $c_1 = 2 \times 15.0912$	$\frac{n_1(t)}{(a_1 \times n_1^3(t) - b_1 \times n_1^2(t) + c_1 \times n_1(t))/3600}$
R_2	$a_2 = 2 \times 1.4877 \times 10^{-7}$, $b_2 = 3 \times 2.9815 \times 10^{-3}$, $c_2 = 2.1 \times 15.0912$	$\frac{n_2(t)}{(a_2 \times n_2^3(t) - b_2 \times n_2^2(t) + c_2 \times n_2(t))/3600}$
R_3	$a_3 = 2 \times 1.4877 \times 10^{-7}$, $b_3 = 2 \times 2.9815 \times 10^{-3}$, $c_3 = 2 \times 15.0912$	$\frac{n_3(t)}{(a_3 \times n_3^3(t) - b_3 \times n_3^2(t) + c_3 \times n_3(t))/3600}$
R_4	$a_4 = 2 \times 1.4877 \times 10^{-7}$, $b_4 = 2 \times 2.9815 \times 10^{-3}$, $c_4 = 2 \times 15.0912$	$\frac{n_4(t)}{(a_4 \times n_4^3(t) - b_4 \times n_4^2(t) + c_4 \times n_4(t))/3600}$

The DUE is solved with an early/late arrival penalty defined in (17). Figure 7(a) depicts the path inflow profiles against the corresponding generalized travel cost profiles and additional costs, i.e., the externality induced by control and state constraints while Figure 7(b) illustrates the trip completion rate profiles against these two costs. As indicated by Figure 7(a), no traveler departs because the travel costs are higher than the minimum travel cost at the very beginning. After a short while, the costs of the two paths attain the DUE cost, and then both paths become active. The departure rates of the two paths keep adjusting to maintain the DUE condition. Since the inflow rate is bounded by the region capacity, additional costs arise between 123 times and 198 times during which the Lagrange multipliers associated with the control constraints are activated. The path travel cost of Path 2 incurs a slight increase at around 208 times and continues to 237 times when the equilibrium condition is broken. The path inflow rates decrease significantly to zero whenever the travel costs of the path are greater than the minimum travel cost, which is consistent with the DUE condition. When the travel cost reaches the equilibrium cost again, travelers choose Path 2 to depart again. From Figure 7(b), travelers' trip completion rate profiles are generally distributed centered along the expected arrival time interval to avoid the schedule delay and penalty cost. The additional cost induced by state and input constraints prevent from boundary queues and over-saturation without introducing complicated boundary queue dynamics as pointed out in Remark 3.2. In this way, travelers achieve minimum generalized travel cost by adjusting their route/path choice and departure time choice, i.e., changing the spatial-temporal distribution of travel demand. Figure 7(c) and Figure 7(d) demonstrates the accumulation state and cumulative outflows of each path. As shown in these figures, flow conservation is well satisfied.

5.3. A network with overlapping paths

Consider a network with overlapping paths adopted from Mariotte and Leclercq (2019) as shown in Figure 8. Two OD pairs are connected by two paths, $R_1 \to R_3 \to R_4 \to R_5 \to R_7$ (Path 1) and $R_2 \to R_3 \to R_4 \to R_6 \to R_8$ (Path 2), respectively. It is assumed that each region admits a well-defined MFD of the form $(G_i(n_i) = a_i \times 1.4877 \times 10^{-7} \times n^3 - b_i \times 2.9815 \times 10^{-3} \times n^2 + c_i \times 15.0912 \times n)/1800$ Detailed regional MFD configurations are listed in Table 2. Total OD travel demands are $[Q_{17}, Q_{28}] = [8000, 6000]$ (units). The planning horizon is $[0, 800]$ (unit-times) while the perferable arrival time interval is $[t_{de}, t_{dl}] = [400, 600]$ (times). A schedule delay penalty of the form (17) is considered. The inflow is restricted to 25 (unit/time) while the accumulation states are restricted to 3400 (units).



(a) Inflow rate and travel cost

(b) Trip completion rate and travel cost

(c) Accumulation state

(d) Cumulative completed trips and travel cost

Figure 7: An example of the Braess like network with nonlinear travel time function

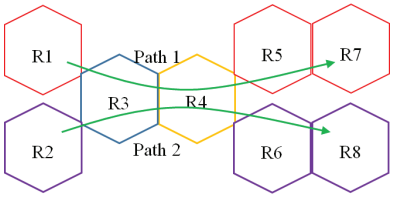
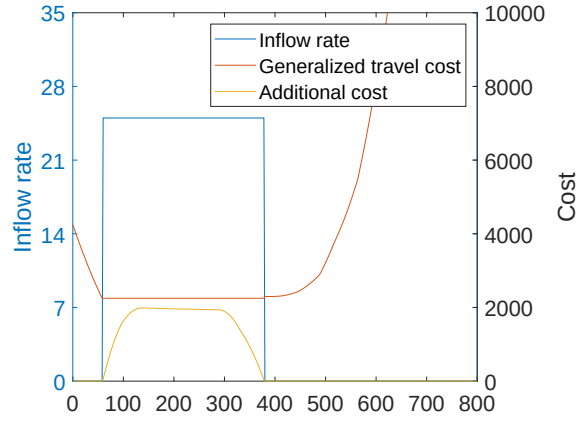


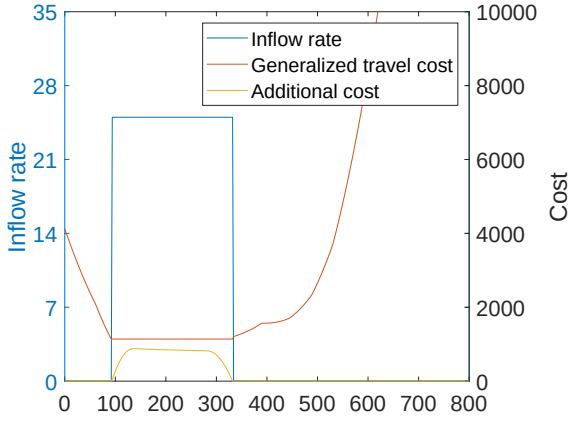
Figure 8: A network with overlapping paths

Table 2: Network configurations: Nonlinear network travel time function

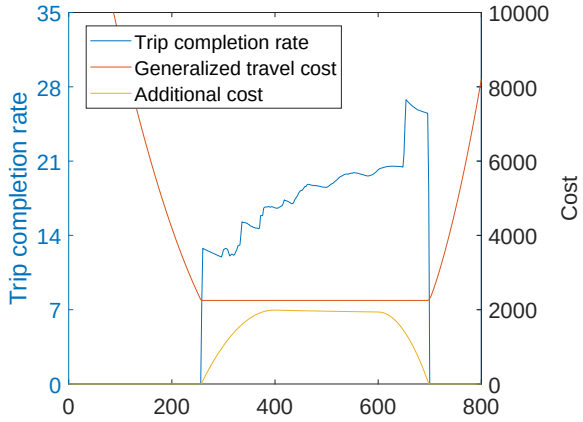
	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
a	3.6	3.6	3.6	3	3.6	3.6	1.8	1.8
b	3.6	3.6	3.6	3	3.6	3.6	1.8	1.8
c	3.6	3.78	3.6	3.15	3.78	3.6	1.89	1.8



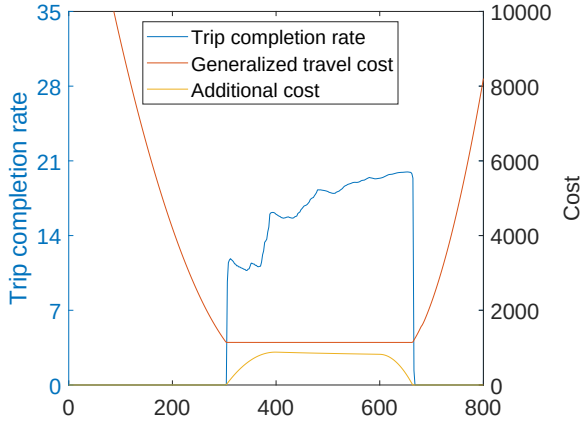
(a) Inflow rate and travel cost: Path 1



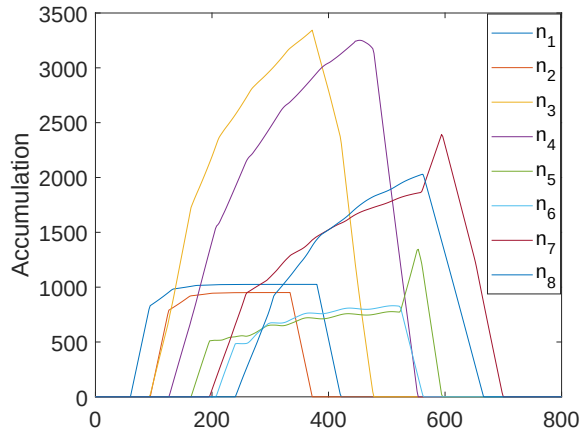
(b) Inflow rate and travel cost: Path 2



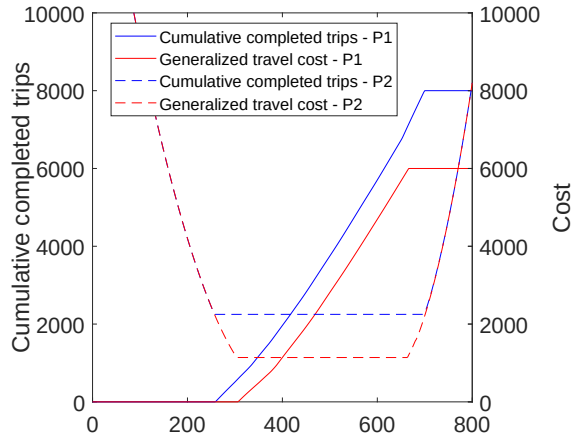
(c) Trip completion rate and travel cost: Path 1



(d) Trip completion rate and travel cost: Path 2



(e) Accumulation state of each region



(f) Cumulative completed trips and travel cost

Figure 9: An example of a network with overlapping paths and nonlinear travel time function

Solving the DUE, the inflow profiles of two paths are depicted against the corresponding generalized travel cost profiles and additional costs in Figure 9(a) and Figure 9(b) respectively. Note there is no route choice but departure time choice within each OD pair. As indicated by Figure 9(a)-Figure 9(b), no traveler in each OD pair departs at the beginning because the early arrival penalty dominates the generalized total cost. During the departure window, the generalized cost stays flat, well satisfying the equilibrium condition. From the behavior realism of travelers, Figure 9(c)-Figure 9(d) indicate that travelers try to finish their trips along the generalized preferable arrival window defined by the DUE condition to minimize their travel cost. In particular, when there is inflow capacity constraint, not all travel demand can be well served during the preferable arrival window. More travelers prefer to arrive early rather than being late (indicated by a larger early arrival time window than the late arrival time window). This is because the consequence of late arrival is more serious in reality. According to Figure 9(e), the network accumulations of all regions are under the corresponding critical points. This verifies the effectiveness of the inflow capacity and accumulation constraints to prevent the network from congestion. Figure 9(f) demonstrates the flow conservation within each OD pair.

6. Conclusions

The MFD framework has been widely used for aggregate modeling of urban traffic dynamics for traffic control, route guidance, parking management and dynamic pricing. Note that neither departure time choice nor route choice under the MFD framework is systematically addressed in the literature. On the other hand, as revealed in recent studies, the MFD model ignoring the travel time to traverse the urban region renders it not preferable as network loading model to investigate departure time choice and route choice behavior. However, when the time-varying travel time is explicitly incorporated in the MFD model for flow propagation, it is very difficult to pursue either analytical or numerical solutions. Nevertheless, the FIFO condition is not always guaranteed when using a nonlinear travel time function.

To tackle the above barriers, this paper proposed a dynamic user equilibrium model with simultaneous departure time and route choices for general urban traffic networks. In line with the literature, an urban traffic network is divided into several regions with each admits a well-defined MFD and thus modeled as multi-region MFD systems. The regional time-varying travel time is explicitly incorporated in the MFD model for proper traffic flow propagation. The DUE problems with both inflow capacity and accumulation constraints were formulated using the DVI framework by Friesz and Han (2018). Dynamic equilibrium conditions were derived analytically under the umbrella of the Pontryagin minimum principle. The proposed DUE gives rise to an alternative approach to capturing the hypercongestion represented by the downward sloping part of the MFD without actually activating traffic congestion while maintaining the FIFO condition using nonlinear travel time functions. A time discretization scheme was devised for numerical simulation of the proposed DUE models. Numerical schemes for calculating various dynamic external costs were proposed by extending the case of linear travel time function proposed in the DTA literature to the general case of nonlinear travel time function under the MFD framework. Numerical examples were conducted to illustrate the characteristics of the DUE while validating the proposed theoretical results. However, the focus of this paper is on the properties of the DUE under the MFD with time delay framework, rather than on the numerical treatments.

For future work, it is interesting to apply the proposed framework for dynamic pricing, vehicle relocation in car-sharing systems and vehicle dispatching in taxi and other on-demand service systems. As a common bottleneck in the DTA and other dynamic optimization problems such as dynamic vehicle routing, an efficient numerical scheme is an urgent need. The approximate dynamic programming (ADP) and

reinforcement learning (RL) provide a flexible framework for solving large-scale dynamic optimization problems.

Acknowledgments

Financial support from the National Key R&D Program of China (Nos. 2018YFB16005), the National Natural Science Foundation of China (Nos. 61573385 & 71501009), the Research Grants Council of Hong Kong (Nos. R5029-18, 15210117E & 15211518E), and the CCF-DiDi Big-Data Joint Lab is gratefully acknowledged.

References

Aboudolas, K. and Geroliminis, N., 2013. Perimeter and boundary flow control in multi-reservoir heterogeneous networks. *Transportation Research Part B: Methodological*, 55:265–281.

Aghamohammadi, R. and Laval, J. A., 2018. Dynamic traffic assignment using the macroscopic fundamental diagram: A review of vehicular and pedestrian flow models. *Transportation Research Part B: Methodological*.

Amirgholy, M. and Gao, H. O., 2017. Modeling the dynamics of congestion in large urban networks using the macroscopic fundamental diagram: User equilibrium, system optimum, and pricing strategies. *Transportation Research Part B: Methodological*, 104:215–237.

Ampountolas, K., Zheng, N., and Geroliminis, N., 2017. Macroscopic modelling and robust control of bi-modal multi-region urban road networks. *Transportation Research Part B: Methodological*, 104:616–637.

Arnott, R., 2013. A bathtub model of downtown traffic congestion. *Journal of Urban Economics*, 76(4):110–121.

Arnott, R. and Buli, J., 2018. Solving for equilibrium in the basic bathtub model. *Transportation Research Part B: Methodological*, 109:150–175.

Arnott, R., Kokoza, A., and Naji, M., 2016. Equilibrium traffic dynamics in a bathtub model: A special case. *Economics of transportation*, 7:38–52.

Buisson, C. and Ladier, C., 2009. Exploring the Impact of Homogeneity of Traffic Measurements on the Existence of Macroscopic Fundamental Diagrams. *Transportation Research Record*, (2124):127–136.

Carey, M. and Mccartney, M., 2002. Behaviour of a whole-link travel time model used in dynamic traffic assignment. *Transportation Research Part B: Methodological*, 36(1):83–95.

Carey, M., Humphreys, P., McHugh, M., and McIvor, R., 2014. Extending travel-time based models for dynamic network loading and assignment, to achieve adherence to first-in-first-out and link capacities. *Transportation Research Part B: Methodological*, 65:90–104.

Daganzo, C. F., 1995. Properties of link travel time functions under dynamic loads. *Transportation Research Part B: Methodological*, 29(2):95–98.

Daganzo, C. F., 2007. Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B: Methodological*, 41(1):49–62.

Daganzo, C. F. and Geroliminis, N., 2008. An analytical approximation for the macroscopic fundamental diagram of urban traffic. *Transportation Research Part B: Methodological*, 42(9):771–781.

Fosgerau, M., 2015. Congestion in the bathtub. *Economics of Transportation*, 4(4):241–255.

Friesz, T. L. and Han, K., 2018. The mathematical foundations of dynamic user equilibrium. *Transportation Research Part B: Methodological, in press*.

Friesz, T. L. and Mookherjee, R., 2006. Solving the Dynamic Network User Equilibrium Problem with State-Dependent Time Shifts. *Transportation Research Part B: Methodological*, 40(3):207–229.

Friesz, T. L., Bernstein, D., Smith, T. E., Tobin, R. L., and Wie, B. W., 1993. A variational inequality formulation of the dynamic network user equilibrium problem. *Operations Research*, 41(1):179–191.

Friesz, T. L., Bernstein, D., Suo, Z., and Tobin, R. L., 2001. Dynamic Network User Equilibrium with State-Dependent Time Lags. *Networks and Spatial Economics*, 1(3):319–347.

Geroliminis, N. and Daganzo, C. F., 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42(9):759–770.

Geroliminis, N. and Sun, J., 2011. Hysteresis phenomena of a macroscopic fundamental diagram in freeway networks. *Transportation Research Part A: Policy and Practice*, 45(9):966–979.

Geroliminis, N., Haddad, J., and Ramezani, M., 2013. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach. *IEEE Transactions on Intelligent Transportation Systems*, 14(1):348–359.

Haddad, J., 2017a. Optimal coupled and decoupled perimeter control in one-region cities. *Control Engineering Practice*, 61: 134–148.

Haddad, J., 2017b. Optimal perimeter control synthesis for two urban regions with aggregate boundary queue dynamics. *Transportation Research Part B: Methodological*, 96:1–25.

Haddad, J. and Mirkin, B., 2016. Adaptive perimeter traffic control of urban road networks based on MFD model with time delays. *International Journal of Robust and Nonlinear Control*, 26(6):1267–1285.

Haddad, J. and Zheng, Z., 2018. Adaptive perimeter control for multi-region accumulation-based models with state delays. *Transportation Research Part B: Methodological, in press*.

Haddad, J., Ramezani, M., and Geroliminis, N., 2013. Cooperative traffic control of a mixed network with two urban regions and a freeway. *Transportation Research Part B: Methodological*, 54:17–36.

Keyvan-Ekbatani, M., Kouvelas, A., Papamichail, I., and Papageorgiou, M., 2012. Exploiting the fundamental diagram of urban networks for feedback-based gating. *Transportation Research Part B: Methodological*, 46(10):1393–1403.

Keyvan-Ekbatani, M., Papageorgiou, M., and Knoop, V. L., 2015a. Controller design for gating traffic control in presence of time-delay in urban road networks. *Transportation Research Part C: Emerging Technologies*, 59:308–322.

Keyvan-Ekbatani, M., Yildirimoglu, M., Geroliminis, N., and Papageorgiou, M., 2015b. Multiple Concentric Gating Traffic Control in Large-Scale Urban Networks. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):2141–2154.

Knoop, V., Hoogendoorn, S., and Van Lint, J., 2012. Routing strategies based on macroscopic fundamental diagram. *Transportation Research Record*, (2315):1–10.

Kouvelas, A., Saeedmanesh, M., and Geroliminis, N., 2017. Enhancing model-based feedback perimeter control with data-driven online adaptive optimization. *Transportation Research Part B: Methodological*, 96:26–45.

Lamotte, R. and Geroliminis, N., 2018. The morning commute in urban areas with heterogeneous trip lengths. *Transportation Research Part B: Methodological*, 117:794 – 810.

Laval, J. A. and Castrillón, F., 2015. Stochastic Approximations for the Macroscopic Fundamental Diagram of Urban Networks. *Transportation Research Part B: Methodological*, 81:904–916.

Laval, J. A., Leclercq, L., and Chiabaut, N., 2018. Minimal parameter formulations of the dynamic user equilibrium using macroscopic urban models: Freeway vs city streets revisited. *Transportation Research Part B: Methodological*, 117:676–686.

Mahmassani, H. S., Saberi, M., and Zockaie, A., 2013. Urban network gridlock: Theory, characteristics, and dynamics. *Transportation Research Part C: Emerging Technologies*, 36:480–497.

Mariotte, G. and Leclercq, L., 2019. Flow exchanges in multi-reservoir systems with spillbacks. *Transportation Research Part B: Methodological*, 122:327–349.

Mariotte, G., Leclercq, L., and Laval, J. A., 2017. Macroscopic urban dynamics: Analytical and numerical comparisons of existing models. *Transportation Research Part B: Methodological*, 101:245–267.

Mirkin, B., Haddad, J., and Shtessel, Y., 2016. Tracking with asymptotic sliding mode and adaptive input delay effect compensation of nonlinearly perturbed delayed systems applied to traffic feedback control. *International Journal of Control*, (9):1–20.

Nie, X. and Zhang, H. M., 2005. Delay-function-based link models: their properties and computational issues. *Transportation Research Part B: Methodological*, 39(8):729–751.

Ramezani, M. and Nourinejad, M., 2018. Dynamic modeling and control of taxi services in large-scale urban networks: A macroscopic approach. *Transportation Research Part C: Emerging Technologies*, 94:203 – 219.

Saberi, M., Mahmassani, H. S., Hou, T., and Zockaie, A., 2014. Estimating Network Fundamental Diagram Using Three-Dimensional Vehicle Trajectories: Extending Edie's Definitions of Traffic Flow Variables to Networks. *Transportation Research Record*, (2422):12–20.

Xu, Y., Wu, J. H., Florian, M., Marcotte, P., and Zhu, D., 1999. Advances in the continuous dynamic network loading problem. *Transportation Science*, 33(4):341–353.

Yang, H. and Huang, H. J., 2005. *Mathematical and economic theory of road pricing*. Elsevier, Amsterdam.

Yildirimoglu, M. and Geroliminis, N., 2014. Approximating dynamic equilibrium conditions with macroscopic fundamental diagrams. *Transportation Research Part B: Methodological*, 70:186–200.

Yildirimoglu, M., Ramezani, M., and Geroliminis, N., 2015. Equilibrium analysis and route guidance in large-scale networks with MFD dynamics. *Transportation Research Part C: Emerging Technologies*, 59:404–420.

Yildirimoglu, M., Sirmatel, I. I., and Geroliminis, N., 2018. Hierarchical control of heterogeneous large-scale urban road networks via path assignment and regional route guidance. *Transportation Research Part B: Methodological*, 118:106–123.

Zheng, N. and Geroliminis, N., 2016. Modeling and optimization of multimodal urban networks with limited parking and dynamic pricing. *Transportation Research Part B: Methodological*, 83:36–58.

Zhong, R., 2011. *Dynamic Assignment, Surveillance and Control for Traffic Network with Uncertainties*. PhD thesis, The Hong Kong Polytechnic University.

Zhong, R., Sumalee, A., Friesz, T. L., and Lam, W. H. K., 2011. Dynamic user equilibrium with side constraints for a traffic network: Theoretical development and numerical solution algorithm. *Transportation Research Part B: Methodological*, 45(7): 1035–1061.

Zhong, R., Sumalee, A., and Maruyama, T., 2012. Dynamic marginal cost, access control, and pollution charge: a comparison of bottleneck and whole link models. *Journal of Advanced Transportation*, 46(3):191–221.

Zhong, R., Chen, C., Huang, Y., Sumalee, A., Lam, W., and Xu, D., 2018a. Robust perimeter control for two urban regions with macroscopic fundamental diagrams: A control-lyapunov function approach. *Transportation Research Part B: Methodological*, 117:687–707.

Zhong, R., Huang, Y., Chen, C., Lam, W., Xu, D., and Sumalee, A., 2018b. Boundary conditions and behavior of the macroscopic fundamental diagram based network traffic dynamics: A control systems perspective. *Transportation Research Part B: Methodological*, 111:327 – 355.

Zhu, D. and Marcotte, P., 2000. On the Existence of Solutions to the Dynamic User Equilibrium Problem. *Transportation Science*, 34(4):402–414.

Appendix A. Nomenclature

We summarize the nomenclature used in this paper in Table A.3.

Appendix B. Comparison of the MFD dynamics and the whole link model

For convenience and comparison, this section presents the differences between the MFD model with time delay and the WLM in the DTA literature. Although the flow propagation of the MFD model with time delay (6) in Section 2) is similar to the WLM, there are several key differences between these two models, e.g., model dynamics and travel time function, as summarized in Table B.4. For example, the travel time function in the WLM is usually specified as a linear function of the link traffic volume $x_a(t)$, i.e.,

$$\tau_a(t) = \tau_0 + \frac{x_a(t)}{C_a}, \tag{B.1}$$

where τ_0 is a flow-invariant travel time (i.e., free-flow travel time) of link a. A physical interpretation of the parameter C_a is the maximum feasible constant outflow from the link, i.e., the service rate of link a (Carey and Mccartney, 2002). Comparing (B.1) and (1), we can see that the travel time function of the WLM depends linearly on the link traffic volume and the constant service rate of link a, which is given exogenously. In contrast, the counterpart of the MFD system with time delay depends on the network output function $G(n(t))$, which is nonlinear and relies on the real-time traffic state of the network endogenously. Despite these differences, it is still possible to analyze the DUE conditions for general traffic networks modeled by the MFD systems with time-varying delay by extending the framework of dynamic traffic assignment with side constraints proposed by Zhong et al. (2011, 2012).

Appendix C. Proof of Proposition 3.1

Proof of Proposition 3.1. In this proof, we derive the necessary condition for the **DUE I** by Pontryagin Minimum Principle. Define

$$\tilde{G}^p_{R_{p,i}}(t) = G^p_{R_{p,i}}\left(t + h_{R_{p,i}}\left(n_{R_{p,i}}(t)\right)\right), \quad \tau^p_{R_{p,i}} = \tau^p_{R_{p,i-1}} + h_{R_{p,i}}\left(n^p_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right)\right), \quad \forall p \in \mathcal{P},\ i \in [1,\, m(p)] \tag{C.1}$$

where $\tau^p_{R_{p,i-1}}$ denotes the time stamp of flow $G^p_{R_{p,i-1}}$ exiting from $R_{p,i-1} \in \mathcal{R}_p$ and entering $R_{p,i}$, and $\tau^p_{R_{p,i}}$ denotes the time stamp of flow $G^p_{R_{p,i}}$ exiting from $R_{p,i}$, which is a recapitulation of (8). The Hamiltonian

Table A.3: Nomenclature

$\mathcal{R}$	The set of all regions
$\mathcal{W}$	The set of all OD pairs
w	OD pair w, $w \in \mathcal{W}$
Q_w	Travel demand of OD pair w
$\mathcal{P}$	The set of all paths, with the number of $\mathcal{P}$ denoted as $\lvert\mathcal{P}\rvert$
p	Path p, $p \in \mathcal{P}$
$\mathcal{P}_w$	The set of all paths connecting OD pair w
$\mathcal{R}_p$	The set of all regions along path p, $\mathcal{R}_p \subseteq \mathcal{R}$
$m(p)$	The number of regions along path p, i.e., $m(p) = \lvert\mathcal{R}_p\rvert$
$m(R)$	The number of paths traversing region R
$q^p(t)$	Inflow rate of path p at time t
R_j	Region R_j in $\mathcal{R}$; the subscript j is omitted when referring to a single region
$n_{R_j}(t)$	Traffic accumulation state of region $R_j \in \mathcal{R}$ at time t
$G_{R_j}(t)$	Total traffic flow (rate) exiting region R_j at time t
$R_{p,i}$	The i^{th} region along path p, $i \in [1, m(p)]$. The relation between the elements of $\mathcal{R}_p$ and $\mathcal{R}$ is given by: If the i^{th} region along path p, i.e., $R_{p,i}$, is the region $R_j \in \mathcal{R}$, then $\delta_{R_{p,i},R_j} = 1$. Otherwise, $\delta_{R_{p,i},R_j} = 0$, i.e., $\delta_{R_{p,i},R_j} = \begin{cases} 1, & \text{if } R_{p,i} = R_j \\ 0, & \text{otherwise} \end{cases}$
$n^p_{R_{p,i}}(t)$	Traffic accumulation state of the i^{th} region along path p at time t. The relation between $n_{R_j}(t)$ and $n^p_{R_{p,i}}(t)$ is given as $n_{R_j}(t) = \sum_{p\in\mathcal{P}} \sum_{i=1}^{m(p)} n_{R_{p,i}}(t)\, \delta_{R_{p,i},R_j}$
$G^p_{R_{p,i}}(t)$	Path-based flow (rate) exiting the i^{th} region along path p, i.e., $R_{p,i}$, at time t. The relationship between $G_{R_j}(t)$ and $G^p_{R_{p,i}}(t)$ is given as $G_{R_j}(t) = \sum_{p\in\mathcal{P}} \sum_{i=1}^{m(p)} G^p_{R_{p,i}}(t)\, \delta_{R_{p,i},R_j}$
$n^{cr}_{R_j}$	The critical accumulation state of region R_j
$h_R(\cdot)$	Travel time (or delay) of region R with $h_R(n_R)\big\rvert_{n_R\to 0}$ denoting the free-flow travel time
$h^p(\cdot)$	Journey time (or path delay operator) of path p
$\tau^p_{R_{p,i}}(t)$	Exit time from the i^{th} region along path p, i.e., $R_{p,i}$, of vehicles departing from the origin at time t
$\Psi^p(\cdot)$	Total travel cost of path p including schedule delay penalty (or effective path delay operator)
ϕ_w	Minimum travel cost of OD pair w
MFD	Macroscopic fundamental diagram
DUE	Dynamic user equilibrium
DTA	Dynamic traffic assignmnet
DVI	Differential variational inequility
FIFO	First-In-First-Out
WLM	Whole link model

Table B.4: Differences between the WLM and the MFD

	MFD system	Whole link model
Network dynamics	Network flow accumulation: regional flow conservation according to endogenous travel time function and a well-calibrated MFD, exogenously bounded by jam accumulation	Link traffic volume: link flow conservation according to pre-specified (or exogenous) travel time function, no exogenous restriction on the volume
Flow propagation	The MFD determines the network outflow rate according to the dynamic network state; the inflow rate and outflow rate are bounded by the network capcity	The link service rate (or capacity) is constant, which is independent of network dynamics; no restriction on the inflow or outflow rate
Queue	The queue is of physical size and gridlock is possible	The queue is vertical and without physical size
Delay	Endogenous: nonlinear depending on the productivity function	Exogenous: linear with respect to the link volume and the exogenous link capacity

function for the optimal control problem is

$$
\begin{aligned}
H_I(t) = & \sum_{w\in\mathcal{W}} \sum_{\forall p\in\mathcal{P}_w} \Bigg\{ \Psi^p(t,\boldsymbol{n}^*)\, q^p(t) + \alpha^p_{R_{p,1}}(t)\Big(q^p(t) - G^p_{R_{p,1}}(t)\Big) + \sum_{i\in[2,\,m(p)]} \alpha^p_{R_{p,i}}(t)\Big(G^p_{R_{p,i-1}}(t) - G^p_{R_{p,i}}(t)\Big) \\
& + \beta^p_{R_{p,1}}(t)\left[q^p(t) - \tilde{G}^p_{R_{p,1}}(t)\Big(1 + h'_{R_{p,1}}\Big(n_{R_{p,1}}(t)\Big)\cdot \dot{n}_{R_{p,1}}(t)\Big)\right] \\
& + \sum_{i\in[2,\,m(p)]} \beta^p_{R_{p,i}}(t)\left[G^p_{R_{p,i-1}}(t) - \tilde{G}^p_{R_{p,i}}(t)\Big(1 + h'_{R_{p,i}}\Big(n_{R_{p,i}}(t)\Big)\cdot \dot{n}_{R_{p,i}}(t)\Big)\right] \\
& - \gamma^p_{R_{p,0}}(t)\, q^p(t) - \sum_{i\in[1,\,m(p)]} \gamma^p_{R_{p,i}}(t)\, G^p_{R_{p,i}}(t) + \sum_{i\in[1,\,m(p)]} \lambda^p_{R_{p,i}}(t)\, n^p_{R_{p,i}}(t) \Bigg\} \\
& + \sum_{R_j\in\mathcal{R}} \zeta_{R_j}(t)\left[\sum_{p\in\mathcal{P}}\Big(q^p(t)\delta_{R_{p,1},R_j} + \sum_{i\in[2,m(p)]} G^p_{R_{p,i-1}}(t)\delta_{R_{p,i},R_j}\Big) - G^{max}_{R_j}\right] + \rho_w(t)\Big(\sum_{p\in\mathcal{P}_w} q^p(t)\Big)
\end{aligned}
\tag{C.2}
$$

The partial derivatives of the Hamiltonian function are

$$
\begin{aligned}
\frac{\partial H_I(t)}{\partial q^p(t)} &= \Psi^p(t,\boldsymbol{n}^*) + \alpha^p_{R_{p,1}}(t) + \beta^p_{R_{p,1}}(t) - \gamma^p_{R_{p,0}}(t) + \sum_{R_j\in\mathcal{R}} \zeta_{R_j}(t)\delta_{R_{p,1},R_j} + \rho_w(t) \\
&= \Psi^p(t,\boldsymbol{n}^*) + \alpha^p_{R_{p,1}}(t) + \beta^p_{R_{p,1}}(t) - \gamma^p_{R_{p,0}}(t) + \zeta_{R_{p,1}}(t) + \rho_w(t),\ \forall\ p\in\mathcal{P} && \text{(C.3a)} \\
\frac{\partial H_I(t)}{\partial G^p_{R_{p,i}}(t)} &= -\alpha^p_{R_{p,i}}(t) + \alpha^p_{R_{p,i+1}}(t) + \beta^p_{R_{p,i+1}}(t) - \gamma^p_{R_{p,i}}(t) + \sum_{R_j\in\mathcal{R}} \zeta_{R_j}(t)\delta_{R_{p,i+1},R_j} \\
&= -\alpha^p_{R_{p,i}}(t) + \alpha^p_{R_{p,i+1}}(t) + \beta^p_{R_{p,i+1}}(t) - \gamma^p_{R_{p,i}}(t) + \zeta_{R_{p,i+1}}(t),\quad \forall p\in\mathcal{P},\quad i\in[1,\,m(p)-1] && \text{(C.3b)} \\
\frac{\partial H_I(t)}{\partial G^p_{R_{p,m(p)}}(t)} &= -\alpha^p_{R_{p,m(p)}}(t) - \gamma^p_{R_{p,m(p)}} && \text{(C.3c)} \\
\frac{\partial H_I(t)}{\partial \tilde{G}^p_{R_{p,i}}(t)} &= -\beta^p_{R_{p,i}}(t)\Big(1 + h'_{R_{p,i}}(n_{R_{p,i}})\cdot \dot{n}_{R_{p,i}}\Big),\ \forall p\in\mathcal{P},\ i\in[1,\,m(p)] && \text{(C.3d)} \\
\frac{\partial H_I(t)}{\partial n^p_{R_{p,i}}(t)} &= \lambda^p_{R_{p,i}}(t),\ \forall p\in\mathcal{P},\ i\in[1,\,m(p)] && \text{(C.3e)}
\end{aligned}
$$

Complementary slackness conditions for nonnegative flow constraints and saturation constraints (13f)-(15) are:

$$q^p(t) \geq 0,\ \gamma^p_{R_{p,0}}(t) \geq 0,\ \gamma^p_{R_{p,0}}(t)\, q^p(t) = 0,\ \forall p \in \mathcal{P} \tag{C.4a}$$

$$G^p_{R_{p,i}}(t) \geq 0,\ \gamma^p_{R_{p,i}}(t) \geq 0,\ \gamma^p_{R_{p,i}}(t)\, G^p_{R_{p,i}}(t) = 0,\ \forall p \in \mathcal{P},\ \forall i \in [1,\, m(p)] \tag{C.4b}$$

$$n^p_{R_{p,i}}(t) \geq 0,\ \lambda^p_{R_{p,i}}(t) \geq 0,\ \lambda^p_{R_{p,i}}\, n^p_{R_{p,i}}(t) = 0,\ \forall p \in \mathcal{P},\ \forall i \in [1,\, m(p)] \tag{C.4c}$$

For an open path p, $\mathcal{R}_p = \{R_{p,1}, R_{p,2}, \cdots, R_{p,i-1}, R_{p,i}, R_{p,i+1}, \cdots, R_{p,m(p)}\}$, with saturation constraint, we have

$$\begin{aligned} G^p_{R_{p,0}}(\tau^p_{R_{p,0}}) &= q^p(\tau^p_{p,0}) > 0 \\ G^p_{R_{p,i}}(\tau^p_{R_{p,i-1}}) > 0,\ G^p_{R_{p,i}}(\tau^p_{R_{p,i}}) &> 0,\ \forall i \in [1,\, m(p)] \\ n^p_{R_{p,i}}(t) > 0,\ \forall t \in [\tau^p_{R_{p,i-1}}, \tau^p_{R_{p,i}}],\ & i \in [1,\, m(p)] \end{aligned}$$

The Lagrange multipliers of the complementary slackness conditions (C.4a)-(C.4c) are then relaxed to

$$\gamma^p_{R_{p,0}}(\tau^p_{R_{p,0}}) = 0 \tag{C.5a}$$

$$\gamma^p_{R_{p,i}}(\tau^p_{R_{p,i-1}}) = \gamma^p_{R_{p,i}}(\tau^p_{R_{p,i}}) = 0,\ \forall i \in [1, m(p)] \tag{C.5b}$$

$$\lambda^p_{R_{p,i}}(t) = 0,\ \forall t \in [\tau^p_{R_{p,i-1}}, \tau^p_{R_{p,i}}],\ \forall i \in [1, m(p)] \tag{C.5c}$$

Stationary conditions at optimality are

$$\left.\frac{\partial H_I}{\partial q^p}\right|_{\tau^p_{R_{p,0}}} = 0,\ \forall w \in \mathcal{W},\ p \in \mathcal{P}_w \tag{C.6a}$$

$$\left.\frac{\partial H_I}{\partial G^p_{R_{p,i}}}\right|_{\tau^p_{R_{p,i}}} + \left[\frac{\partial H_I}{\partial \tilde{G}^p_{R_{p,i}}} \frac{1}{1 + h'_{R_{p,i}}(n_{R_{p,i}}) \cdot \dot{n}_{R_{p,i}}}\right]_{\tau^p_{R_{p,i-1}}} = 0,\ \forall p \in \mathcal{P},\ i \in [1,\, m(p)] \tag{C.6b}$$

In conjunction with (C.5c), the adjoint equations give

$$\frac{\mathrm{d}\alpha^p_{R_{p,i}}(t)}{\mathrm{d}t} = -\frac{\partial H_I(t)}{\partial n^p_{R_{p,i}}(t)} = -\lambda^p_{R_{p,i}}(t) = 0,\ \forall p \in \mathcal{P},\ i \in [1, m(p)],\ t \in [\tau^p_{R_{p,i-1}}, \tau^p_{R_{p,i}}] \tag{C.7a}$$

$$\frac{\mathrm{d}\rho_w(t)}{\mathrm{d}t} = 0,\ \forall w \in \mathcal{W} \tag{C.7b}$$

Boundary conditions at terminal time T are

$$\alpha^p_{R_{p,i}}(T) = 0,\ -\rho_w(T) - \phi_w = 0,\ \forall w \in \mathcal{W},\ \forall p \in \mathcal{P}_w,\ i \in [1, m(p)] \tag{C.8}$$

(C.7a) and (C.8) give

$$\alpha^p_{R_{p,i}}(\tau^p_{R_{p,i}}) - \alpha^p_{R_{p,i}}(\tau^p_{R_{p,i-1}}) = 0,\ \forall p \in \mathcal{P},\ i \in [1, m(p)], \tag{C.9}$$

(C.7b) and (C.8) give

$$\rho_w(t) = -\phi_w = \rho_w,\ \forall w \in \mathcal{W} \tag{C.10}$$

In conjunction with (13k), (C.5a) and (C.10), (C.6a) gives

$$\begin{aligned}\left.\frac{\partial H_I}{\partial q^p}\right|_{\tau^p_{R_{p,0}}} &= \Psi_p\left(\tau^p_{R_{p,0}}, \boldsymbol{n}^*\right) + \alpha^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \beta^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) - \gamma^p_{R_{p,0}}\left(\tau^p_{R_{p,0}}\right) + \zeta^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \rho_w\left(\tau^p_{R_{p,0}}\right) \\ &= \Psi_p\left(\tau^p_{R_{p,0}}, \boldsymbol{n}^*\right) + \alpha^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \beta^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \zeta^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \rho_w \\ &= 0, \quad \forall w \in \mathcal{W},\ p \in \mathcal{P}_w\end{aligned} \tag{C.11}$$

where $\rho_w(t) = \rho_w$, is a constant, and $\gamma^p_{R_{p,0}}\left(\tau^p_{R_{p,0}}\right) = 0$ for an open path. In conjunction with (C.3b) and (C.3d), (C.6b) gives

$$\begin{cases} -\alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) + \alpha^p_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) + \beta^p_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) + \zeta_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) - \beta^p_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) = 0, \quad i \in [1,\, m(p)-1] \\ -\alpha^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)}}\right) - \beta^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)-1}}\right) = 0, \end{cases} \quad \forall p \in \mathcal{P}$$

Denote $\tilde{\beta}^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) \doteq \beta^p_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right),\ \forall p \in \mathcal{P},\ i \in [1,\, m(p)]$, then

$$\begin{cases} \alpha^p_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) + \beta^p_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) = \alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) + \tilde{\beta}^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) - \zeta_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right), \quad \forall i \in [1,\, m(p)-1] \\ \beta^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)-1}}\right) = \tilde{\beta}^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)}}\right) = -\alpha^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)}}\right), \end{cases} \quad \forall p \in \mathcal{P} \tag{C.12}$$

Combining (C.9) and (C.12) gives

$$D_{m(p)} = \alpha^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)-1}}\right) + \beta^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)-1}}\right) = \alpha^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)-1}}\right) - \alpha^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)}}\right) = 0$$

and $\forall i \in [1,\, m(p)-1]$, there are

$$\begin{aligned} D_{i+1} &= \alpha^p_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) + \beta^p_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) = \alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) + \tilde{\beta}^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) - \zeta_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) \\ D_i &= \alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) + \beta^p_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) = \alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) + \tilde{\beta}^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) \\ D_{i+1} - D_i &= \alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) - \alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) - \zeta_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) = -\zeta_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) \end{aligned}$$

Proceeding along this stream gives

$$\begin{aligned} D_i &= D_{i+1} + \zeta_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right), \quad \forall\, i \in [1,\, m(p)-1] \\ D_1 &= \alpha^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \beta^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) = D_{m(p)} + \sum_{i=2}^{m(p)} \zeta_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) = \sum_{i=2}^{m(p)} \zeta_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) \end{aligned}$$

Thus

$$l_p(\tau^p_{R_{p,0}}) = \alpha^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \beta^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \zeta_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) = D_1 + \zeta_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) = \sum_{i=1}^{m(p)} \zeta_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) \tag{C.13}$$

Combining (C.11) and (C.13) gives

$$\Psi_p\left(\tau^p_{R_{p,0}}, \boldsymbol{n}^*\right) + \alpha^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \beta^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \zeta^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \rho_w = \Psi_p\left(\tau^p_{R_{p,0}}, \boldsymbol{n}^*\right) + l_p(\tau^p_{R_{p,0}}) + \rho_w = 0,\ \forall w \in \mathcal{W}, p \in \mathcal{P}_w$$

In conjunction with (C.10), we have

$$\Psi^p\left(\tau^p_{R_{p,0}}, \boldsymbol{n}^*\right) + l_p(\tau^p_{R_{p,0}}) = \phi_w,\ p \in \mathcal{P}_w \Rightarrow q^{p*}(\tau^p_{R_{p,0}}) > 0$$

Note that ϕ_w is constant and the $\tau^p_{R_{p,0}}$ is arbitrary, $\tau^p_{R_{p,0}}$ can be restated as t, the DUE conditions with inflow saturation constraint are:

$$\Psi^p\left(t, \boldsymbol{n}^*\right) + l_p(t) = \phi_w,\ p \in \mathcal{P}_w \Rightarrow q^{p*}(t) > 0$$
$$\Psi^p\left(t, \boldsymbol{n}^*\right) + l_p(t) > \phi_w,\ p \in \mathcal{P}_w \Rightarrow q^{p*}(t) = 0$$

Appendix D. Proof of Corollary 3.1

Proof of Corollary 3.1. This section presents the proof of the **DUE II**. Introducing state constraint (15) in the **DUE I** [4], the Hamilton function is extended as:

$$H_{II}(t) = H_I(t) + \sum_{R_j \in \mathcal{R}} \eta_{R_j}(t)\left(\sum_{p\in\mathcal{P}} \sum_{i\in[1,m(p)]} n^p_{R_i}(t)\,\delta_{R_{p,i},R_j} - n^{cr}_{R_j}\right)$$

Partial derivatives (C.3a)-(C.3d), indirectly related to state variable $n^p_{R_{p,i}}(t)$ or (15), stay the same as in Appendix C. Distinct partial derivatives are

$$\frac{\partial H_{II}(t)}{\partial n^p_{R_{p,i}}(t)} = \lambda^p_{R_{p,i}}(t) + \sum_{R_j\in\mathcal{R}} \eta_{R_j}(t)\,\delta_{R_{p,i},R_j} = \lambda^p_{R_{p,i}}(t) + \eta_{R_{p,i}}(t),\ \forall p \in \mathcal{P},\ i \in [1,\, m(p)]$$

Then the adjoint equation gives

$$\frac{\mathrm{d}\alpha^p_{R_{p,i}}(t)}{\mathrm{d}t} = -\frac{\partial H_{II}(t)}{\partial n^p_{R_{p,i}}(t)} = -\lambda^p_{R_{p,i}}(t) - \eta_{R_{p,i}}(t) = -\eta_{R_{p,i}}(t)\ \ \forall p \in \mathcal{P},\ i \in [1, m(p)],\ t \in [\tau^p_{R_{p,i-1}}, \tau^p_{R_{p,i}}]$$

and (C.9) is restated as:

$$\alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) - \alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) = \int_{\tau^p_{R_{p,i-1}}}^{\tau^p_{R_{p,i}}} \eta_{R_{p,i}}(t)\mathrm{d}t \tag{D.1}$$

Following the proceeding stream in Appendix C, combining (D.1) and (C.12) gives

$$D_{m(p)} = \alpha^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)-1}}\right) + \beta^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)-1}}\right) = \alpha^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)-1}}\right) - \alpha^p_{R_{p,m(p)}}\left(\tau^p_{R_{p,m(p)}}\right) = \int_{\tau^p_{R_{p,m(p)-1}}}^{\tau^p_{R_{p,m(p)}}} \eta_{R_{p,m(p)}}(t)\,\mathrm{d}t$$

[4] A verification of constraint qualifications can be obtained by extending the analysis in Zhong et al. (2011) as the state constraint is of the same form as that in Zhong et al. (2011).

and $\forall i \in [1,\, m(p)-1]$, there are

$$\begin{aligned} D_{i+1} &= \alpha^p_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) + \beta^p_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) = \alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) + \tilde{\beta}^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) - \zeta_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) \\ D_i &= \alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) + \beta^p_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) = \alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) + \tilde{\beta}^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) \\ D_{i+1} - D_i &= \alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i}}\right) - \alpha^p_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) - \zeta_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) = -\int_{\tau^p_{R_{p,i-1}}}^{\tau^p_{R_{p,i}}} \eta_{R_{p,i}}(t)\,\mathrm{d}t - \zeta_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) \end{aligned}$$

Proceeding along this stream gives

$$\begin{aligned} D_i &= D_{i+1} + \int_{\tau^p_{R_{p,i-1}}}^{\tau^p_{R_{p,i}}} \eta_{R_{p,i}}(t)\,\mathrm{d}t + \zeta_{R_{p,i+1}}\left(\tau^p_{R_{p,i}}\right) \quad \forall\, i \in [1,\, m(p)-1] \\ D_1 &= \alpha^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \beta^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) = D_{m(p)} + \sum_{i=1}^{m(p)-1} \int_{\tau^p_{R_{p,i-1}}}^{\tau^p_{R_{p,i}}} \eta_{R_{p,i}}(t)\mathrm{d}t + \sum_{i=2}^{m(p)} \zeta_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) \\ &= \sum_{i=1}^{m(p)} \int_{\tau^p_{R_{p,i-1}}}^{\tau^p_{R_{p,i}}} \eta_{R_{p,i}}(t)\mathrm{d}t + \sum_{i=2}^{m(p)} \zeta_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right) \end{aligned}$$

Thus

$$l_p(\tau^p_{R_{p,0}}) = \alpha^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \beta^p_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) + \zeta_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) = D_1 + \zeta_{R_{p,1}}\left(\tau^p_{R_{p,0}}\right) = \sum_{i=1}^{m(p)} \int_{\tau^p_{R_{p,i-1}}}^{\tau^p_{R_{p,i}}} \eta_{R_{p,i}}(t)\,\mathrm{d}t + \sum_{i=1}^{m(p)} \zeta_{R_{p,i}}\left(\tau^p_{R_{p,i-1}}\right)$$

In line with Appendix C, the DUE conditions with inflow and state saturation constraints are:

$$\begin{aligned} &\Psi^p\left(t, \boldsymbol{n}^*\right) + l_p(t) = \phi_w,\ p \in \mathcal{P}_w \Rightarrow q^{p*}(t) > 0 \\ &\Psi^p\left(t, \boldsymbol{n}^*\right) + l_p(t) > \phi_w,\ p \in \mathcal{P}_w \Rightarrow q^{p*}(t) = 0 \end{aligned}$$

Appendix E. Proof of FIFO

In this section, we show that if $n \le n_{cr}$, the nonlinear travel time function $h(n(t))$ satisfies the FIFO condition. Note that the travel time function $h(n(t))$ as a monotone increasing function of n satisfies [5]

$$h'(n(t)) \ge 0 \tag{E.1}$$

According to Carey et al. (2014), FIFO holds if and only if

$$1 + h'(n(t))\dot{n}(t) \ge 0 \tag{E.2}$$

When $\dot{n}(t) \ge 0$, in line with (E.1), we have $1 + h'(n(t))\dot{n}(t) \ge 1 > 0$, and thus the FIFO holds.

When $\dot{n}(t) < 0$, the maximum traffic dissolve rate can be achieved when there is no inflow (i.e., $q(t) = 0$) and the outflow rate $G(t)$ is at its maximum G_{max}, i.e.,

$$\dot{n}(t) = q(t) - G(n(t)) \ge -\max(G(t))$$

[5] "$'$" denotes differentiation with respect to the associated function argument, while "$\cdot$" denotes differentiation with respect to time t.

The maximum throughput is achieved at the critical point $n_{cr}, 0 < n_{cr} < n_{jam}$, then $G'(n_{cr}) = 0$, where n_{jam} is the jammed accumulation. Using the setting calibrated in Geroliminis and Daganzo (2008), we depict h' in Figure E.10. We define the point at which $h''(\cdot) = 0$ as n_{upper}, such that

$$\begin{cases} h''(n_{upper}) = 0 \\ h''(n) \geq 0, n \leq n_{upper} \\ h''(n) < 0, n > n_{upper} \end{cases}$$

Generally, we have $n_{cr} \leq n_{upper} \leq n_{jam}$.

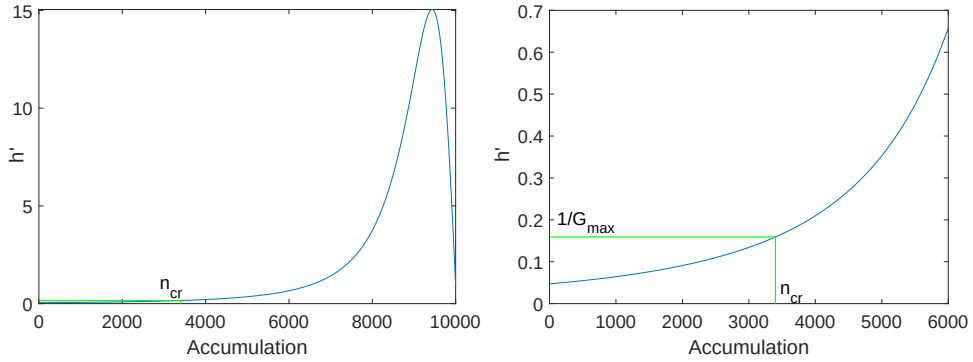


Figure E.10: Differential of h with respect to n

For any $n \in [0, n_{cr}]$, $h''(n) \geq 0$ indicates that $h'(n)$ is monotone increasing with the argument, then

$$\max(h'(n)) = h'(n_{cr}) = \frac{1}{G(n_{cr})} - \frac{n_{cr}G'(n_{cr})}{G^2(n_{cr})} = \frac{1}{G_{max}} - 0 = \frac{1}{G_{max}}$$

$$1 + h'(n)\dot{n} \geq 1 - h'(n)G_{max} \geq 1 - \max(h'(n))G_{max} = 1 - \frac{G_{max}}{G_{max}} = 0$$

Therefore, the FIFO holds if $n \in [0, n_{cr}]$. However, for $n \in (n_{cr}, n_{upper}]$, the FIFO condition cannot always be guaranteed.