Extending the Link Transmission Model with non-triangular fundamental diagrams and capacity drops

van der Gun, Jeroen; Pel, Adam; van Arem, Bart

DOI
10.1016/j.trb.2016.12.011

Publication date
2017

Document Version
Accepted author manuscript

Published in
Transportation Research. Part B: Methodological

Citation (APA)
van der Gun, J., Pel, A., & van Arem, B. (2017). Extending the Link Transmission Model with non-triangular fundamental diagrams and capacity drops. *Transportation Research. Part B: Methodological*, *98*, 154-178. Article 9. https://doi.org/10.1016/j.trb.2016.12.011

Extending the Link Transmission Model with non-triangular fundamental diagrams and capacity drops

Jeroen P.T. van der Gun*, Adam J. Pel, Bart van Arem
Delft University of Technology. P.O. Box 5048, 2600 GA Delft, The Netherlands
* Corresponding author. E-mail: J.P.T.vanderGun@tudelft.nl

Published in
Transportation Research Part B: Methodological
Volume 98, pages 154-178, April 2017, doi:10.1016/j.trb.2016.12.011

Abstract

The original Link Transmission Model as formulated by Yperman et al. (2006) simulates traffic according to Lighthill-Whitham-Richards theory with a very small numerical error, yet only supports triangular fundamental diagrams. This paper relaxes that restriction in two steps. Firstly, we extend the model to handle any continuous concave fundamental diagram, and prove that this extension is still consistent with Lighthill-Whitham-Richards theory. Secondly, we extend the theory and model to handle a capacity drop, explicitly looking into the handling of both the onset and release of congestion. The final model is still first-order and suitable for general networks. Numerical examples show that it qualitatively improves on the original model due to uniquely featuring complex traffic patterns including stop-and-go waves, with crisp shockwaves between traffic states, as well as acceleration fans.

Keywords: Link Transmission Model, Lighthill-Whitham-Richards theory, first-order model, capacity drop, node model, stop-and-go wave.

1. Introduction

Lighthill-Whitham-Richards (LWR) theory or kinematic wave theory, introduced by Lighthill and Whitham (1955) and Richards (1956), consists of two main equations: the conservation of vehicles and the equilibrium flow-density relationship. Assuming that traffic is always in an equilibrium state, these combine into a single partial differential equation for the propagation of traffic along a network link. Traditionally, this partial differential equation has often been solved by the Cell Transmission Model (CTM) (Daganzo, 1994), that discretizes roads into small cells according to the Godunov (1959) scheme. The Lagged Cell Transmission Model (LCTM) (Daganzo, 1999) and its later enhancement (Szeto, 2008) are variants of this method, reducing the numerical error.

Newell (1993) proposed a very different solution scheme, using cumulative numbers of vehicles as the primary variable. Later, this idea led to the development of the Link Transmission Model (LTM) (Yperman et al., 2006; Yperman, 2007), which does not discretize space and consequently leads to substantially smaller numerical errors (or computation time) than both the CTM and the LCTM. Daganzo (2005a; 2005b) and Jin (2015) show implicitly and Han et al. (2016) show explicitly that this numerical procedure indeed solves the partial differential equation as the time step tends to zero, but only for triangular fundamental diagrams (FDs).

However, the requirement of triangular FDs is restrictive in multiple ways. Edie (1961) already identified that the speed in subcritical traffic decreases as traffic density increases. It is important to capture this relationship for the modeling of travel times in light traffic, yet a triangular FD does not do so. A non-linear free-flow branch in the FD furthermore captures platoon dispersion (Geroliminis and Skabardonis, 2005). Edie also recognized that there may be a discontinuity between the free-flow capacity and the queue discharge rate. This is commonly referred to as the capacity drop, i.e. the effect that the presence of congestion reduces the maximum flow. Papageorgiou (1998) mentioned this as an important aspect of traffic flow that models should be able to reproduce, particularly when considering traffic control. This is especially relevant when testing or optimizing traffic control measures aimed at preventing or postponing the occurrence of the capacity drop, like many ramp metering installations, or at dissolving stop-and-go waves or wide moving jams, like SPECIALIST (Hegyi et al., 2008). While Hajiahmadi et al. (2013) propose an extension to the LTM for variable speed limits and ramp metering, the lack of a capacity drop in triangular FDs thus significantly restricts its usability, e.g. in assessing control strategies. Separate modeling of a free-flow capacity and a queue discharge rate can furthermore be expected to benefit strategic assessments of intelligent in-vehicle systems designed to intervene specifically in case of congestion, such as the Congestion Assistant (Van Driel and Van Arem, 2010).

Unlike the continuous-space LTM, these issues have mostly been addressed for the discrete-space models. General continuous FDs can be handled by the CTM and LCTM with proven convergence to LWR theory (Daganzo, 1995; Daganzo, 1999; Szeto, 2008) and have been incorporated into CTM-based optimization problems (Nie, 2011; Carey and Watling, 2012). Multiple different modifications of the CTM have been proposed to deal with a capacity drop (see Section 4). None of this has so far been the case for the LTM.

Hence, the purpose of this paper is to overcome the aforementioned limitations of the shape of the FD in the LTM. More specifically, we extend the LTM to handle general concave FDs, optionally including capacity drops. The resulting model, which we show to converge to LWR theory if there is no capacity drop, is applicable to general networks and features both standing queues, with a head initially fixed at the bottleneck that may move upstream later, and stop-and-go waves that can both grow and dissolve. Qualitative properties of this model are demonstrated with numerical examples.

This paper is structured as follows. First, Section 2 will briefly introduce the structure of the LTM, consisting of a link model and a node model. Then, in Section 3, we derive a link model algorithm for the general case of a continuous concave FD, proving its consistency with LWR theory, and compare it to other link model formulations previously proposed in literature. Next, in Sections 4-6, we review previous work extending first-order models with capacity drops and subsequently extend LWR theory, the previous link model and a node model to allow for a capacity drop. Section 7

demonstrates the final model with two numerical examples. Finally, we list our conclusions in Section 8.

2. Structure of the LTM

The starting point of this paper is the LTM for dynamic network loading, as formulated by Yperman et al. (2006) and Yperman (2007). Its primary components are a link model and a node model, that together are used to update in time steps of size Δt the cumulative number of vehicles $N(x,t)$ at the entrance $x_{i,0}$ and exit $x_{i,L}$ of each link i.

The link model is used to determine the sending and receiving flows, which for each time step indicate the number of vehicles that could potentially exit and enter the link respectively. For link i, these quantities are denoted $S_i(t)$ and $R_i(t)$ respectively. The procedures for determining them are similar and will be discussed in more detail below.

The node model then considers the interactions of traffic at intersections to derive transition flows $G_{ij}(t)$, the number of vehicles that succeed in crossing the intersection, where different node models can be used to represent different types of intersections. They indicate how much of each turn-specific sending flow $S_{ij}(t)$ will actually pass the node during the time step.

The node model also needs to know the turning fractions $S_{ij}(t)/S_i(t)$ as input, which can be either specified exogenously or modeled endogenously by splitting the traffic into multiple so-called commodities with different routing behavior. Yperman et al. (2006) and Yperman (2007) assumed the latter option in their model formulation. Although we do not consider the specification of turning fractions in this paper, the results of this paper can be used in both these cases.

Algorithm 1 below summarizes the overall process, showing how the link model and the node model together specify the traffic flow propagation.

Algorithm 1. Link Transmission Model.

- For each time step t for each node:
  - Using the link model, determine sending flow $S_i(t)$ for each incoming link i.
  - Using the link model, determine receiving flow $R_j(t)$ for each outgoing link j.
  - Determine turning fractions $\dfrac{S_{ij}(t)}{S_i(t)}$ for each turn ij.
  - Using a node model, determine transition flows $G_{ij}(t)$ for each turn ij.
  - $N(x_{i,L},t+\Delta t) := N(x_{i,L},t)+\sum_j G_{ij}(t)$ for each incoming link i.
  - $N(x_{j,0},t+\Delta t) := N(x_{j,0},t)+\sum_i G_{ij}(t)$ for each outgoing link j.

The LTM discretizes only time, not space. Because due to the Courant-Friedrichs-Lewy (1928) condition, the maximum possible time step of a node depends on the length of the attached links, Yperman (2007) suggested that different nodes may be operated with different time step sizes to retain a high computational efficiency without being restricted by the smallest link in a (large)

network. In this paper we explicitly incorporate this suggestion by writing Δt_{x_0} and Δt_{x_L} for the time step sizes of the upstream and downstream nodes of a link respectively and only using Δt without subscript if the node under consideration is obvious.

Note that two close variants of the explicit forward simulation scheme of Algorithm 1 have been proposed recently. Himpe et al. (2016) turned it into an implicit iterative scheme, permitting the use of larger time steps. At least Section 3 of this paper regarding continuous concave FDs is also compatible with this variant. Alternatively, Hajiahmadi et al. (2016), Van de Weg et al. (2016) and Long et al. (2016) turned Algorithm 1 into optimization problems, with points of the cumulative curves as decision variables and the link and node models as constraints. Because for non-triangular FDs we find constraints that only apply conditionally, this paper is likely not compatible with that variant. Further discussion of these two variants is outside the scope of this paper.

3. Link model for continuous concave FDs

In this section, we will define a link model for the case of a continuous concave FD $Q(k)$, i.e. without capacity drop. First, Subsection 3.1 introduces our notation and axioms. Next, we derive a method to compute sending and receiving flows in Subsection 3.2 resulting in the algorithms listed in Subsection 3.3. Finally, Subsection 3.4 compares this newly proposed model with other models in literature.

3.1. Notation and axioms

A general continuous concave FD $Q(k)$ is depicted in Fig. 1 below. Here (k_C, q_C) denotes the capacity point and k_J denotes jam density. Note that we omit link indices on all variables for brevity.

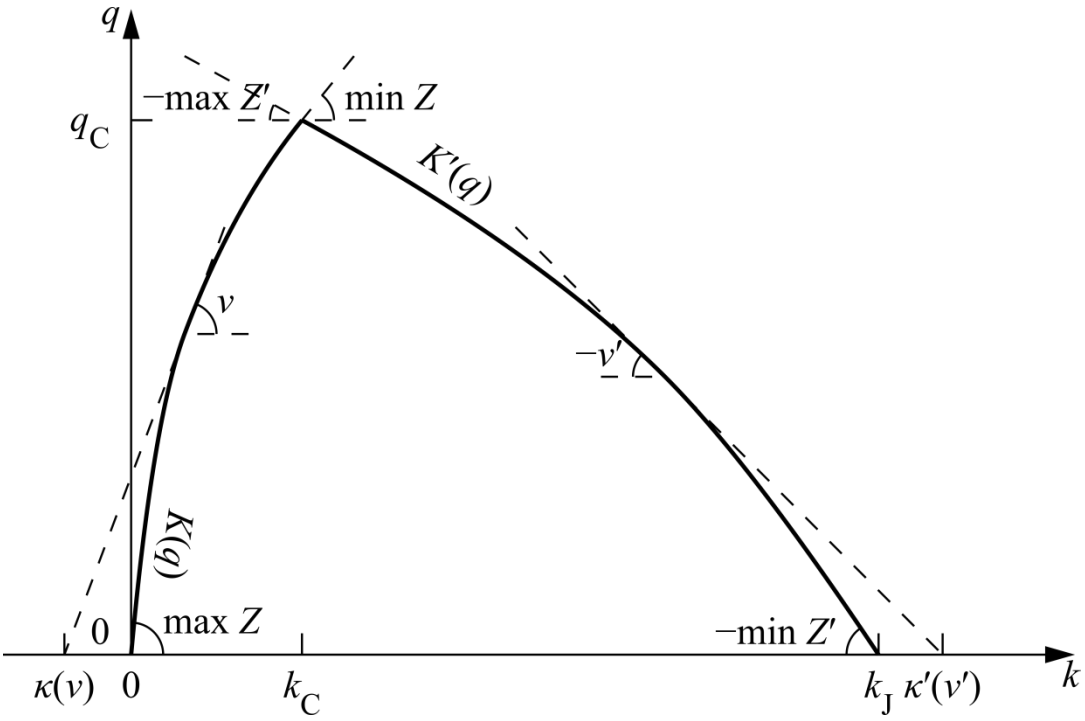

Fig. 1. Example fundamental diagram satisfying Eq. (2).

As an alternative to a function $Q(k)$, the FD may be written as two functions $K(q)$ and $K'(q)$, describing the free-flow branch and the congested branch respectively. For each branch, we define a set of relevant wave speeds, indicating the speeds at which traffic states propagate, as

$$\begin{aligned} Z &= \left[\inf \operatorname{im} \frac{dq}{dK}, \sup \operatorname{im} \frac{dq}{dK} \right] \\ Z' &= \left[\inf \operatorname{im} \frac{dq}{dK'}, \sup \operatorname{im} \frac{dq}{dK'} \right] \end{aligned} . \tag{1}$$

The prime notation indicates the congested branch rather than a derivative and im is used to denote the image of a function. We impose the following restriction on the shape of the FD:

$$\left\{ \min Z, \max Z, \min Z', \max Z' \right\} \cap \left\{ -\infty, 0, \infty \right\} = \varnothing . \tag{2}$$

Note that contrary to this restriction, from a traffic flow theory point of view, $\min Z = 0$ and $\max Z' = 0$ could be permitted, which would allow the FD to be horizontal at the capacity point. However, we will later see that the full restriction allows to compute sending and receiving flows efficiently and exactly on the link level. Note that we do not require the branches of the FD to be continuously differentiable, e.g. piecewise-linear FDs can be used if desired.

Let us now define function $\kappa(v)$ that indicates where a tangent line of slope v to the free-flow branch would intersect the density axis. Likewise, we define function $\kappa'(v')$ for the congested branch. The corresponding formulas, domains and images are

$$\begin{aligned} \kappa(v) &= \min_{q \in \operatorname{dom} K} \left(K(q) - \frac{q}{v} \right), \quad \operatorname{dom} \kappa = (0, \max Z], \quad \operatorname{im} \kappa = (-\infty, 0] \\ \kappa'(v') &= \max_{q \in \operatorname{dom} K'} \left(K'(q) - \frac{q}{v'} \right), \quad \operatorname{dom} \kappa' = [\min Z', 0), \quad \operatorname{im} \kappa' = [k_J, \infty) \end{aligned} . \tag{3}$$

Conversely, by appropriately taking the upper and lower envelopes of these tangent lines, we get back the FD:

$$\begin{aligned} &\forall q \in \operatorname{dom} K: \quad K(q) = \max_{v \in Z} \left(\frac{q}{v} + \kappa(v) \right) \\ &\forall q \in \operatorname{dom} K': \quad K'(q) = \min_{v' \in Z'} \left(\frac{q}{v'} + \kappa'(v') \right) \\ &\forall k \in [0, k_J]: \quad Q(k) = \min \left(\min_{v \in Z} v\left(k - \kappa(v)\right), \min_{v' \in Z'} v'\left(k - \kappa'(v')\right) \right) \end{aligned} . \tag{4}$$

We furthermore define $V(q)$ and $V'(q)$ as the sets of positive and negative tangent line slopes compatible with the FD at flow q:

$$\begin{aligned} V(q) &= \arg\max_{v \in Z} \left(\frac{q}{v} + \kappa(v) \right) = \left\{ v \in Z \,\middle|\, \frac{q}{v} + \kappa(v) = K(q) \right\} \\ V'(q) &= \arg\min_{v' \in Z'} \left(\frac{q}{v'} + \kappa'(v') \right) = \left\{ v' \in Z' \,\middle|\, \frac{q}{v'} + \kappa'(v') = K'(q) \right\} \end{aligned} . \tag{5}$$

The theoretical basis for traffic propagation along the link is formed by LWR theory. Traditionally, this takes the form of the following scalar conservation law, which combines conservation of vehicles with the FD:

$$\frac{\partial k}{\partial t}+\frac{dQ(k)}{dk}\frac{\partial k}{\partial x}=0\,. \tag{6}$$

However, it is more convenient to replace this differential equation with the following Hamilton-Jacobi equation, that also works if $Q(k)$ is not continuously differentiable:

$$\frac{\partial N}{\partial t}-Q\left(-\frac{\partial N}{\partial x}\right)=0,\quad q=\frac{\partial N}{\partial t},\quad k=-\frac{\partial N}{\partial x}\,. \tag{7}$$

This differential equation states the FD in a way that implicitly guarantees conservation of vehicles (Newell, 1993):

$$\frac{\partial k}{\partial t}+\frac{\partial q}{\partial x}=-\frac{\partial^2 N}{\partial x\partial t}+\frac{\partial^2 N}{\partial t\partial x}=0\,. \tag{8}$$

The differential equation is combined with the Lax (1957) shock admissibility or entropy condition to get a unique weak solution, eliminating alternative solutions where acceleration fans are replaced with shocks. This is achieved by prohibiting shocks from emanating waves, thus allowing discontinuities in flow or density only when they absorb waves or run parallel to them[1]. Then, because Hamiltonian $-Q(-\partial N/\partial x)$ is convex in $\partial N/\partial x$, the Hamilton-Jacobi equation, and thus the propagation of traffic along a link, can be solved with relative ease using variational theory (Evans, 2002; Daganzo, 2005a). We use this in the next subsection to determine the sending and receiving flows, assuming that the link is initially empty.

3.2. Computing sending and receiving flows: solution networks

The sending and receiving flows will be solved in terms of cumulative numbers of vehicles. More precisely, our algorithm relies on finding the maximum possible $N(x,t+\Delta t_x)$ at the considered end of the link $x\in\{x_0,x_L\}$ at the end of the time step under consideration, so that $N(x,t+\Delta t_x)-N(x,t)$ is the maximum number of vehicles exiting or entering the link during the time step, which are the sending flow $S(t)$ and the receiving flow $R(t)$ respectively.

We thus rephrased the traffic propagation problem into finding the maximum possible value of $N(x,t+\Delta t_x)$ for $x\in\{x_0,x_L\}$. To do so, we apply the variational theory developed by Daganzo (2005a; 2005b). The boundary condition for this application is formed by the values of the cumulative curves in previous time steps at both link ends. We build a solution network, as defined by Daganzo (2005a), that indicates how each boundary point may constrain the cumulative number of vehicles N at our solution point $P=(x_P,t+\Delta t_{x_P})$. This applies to the determination of both the sending and receiving flow, hence we derive both solution networks simultaneously in the ensuing.

[1] If a discontinuity runs parallel to waves on both sides, it is a contact discontinuity. If it absorbs waves from one or both sides, it is a shockwave.

Fig. 2 illustrates the concept of using solution networks. After each time step, the boundary condition is extended with the newly found solution and the solution network is shifted to compute the next time step. Since we assume the link is initially empty, we can disregard the initial condition by moving it sufficiently far to the left while extending the boundary conditions into the past with $N = 0$.

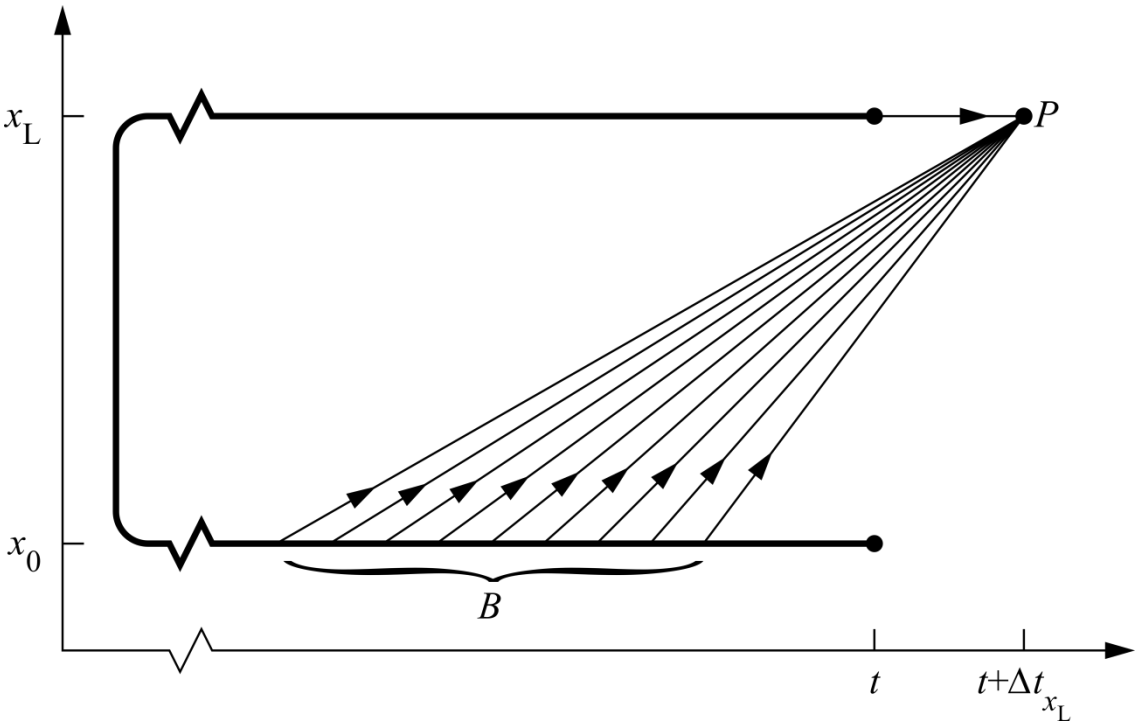

Fig. 2. Concept of a solution network for the sending flow, showing its boundary condition and space-time paths.

Considering what space-time paths to include in our solution network, Fact 3 of Daganzo (2005b) states that we only need to consider straight space-time paths from the boundary to the solution point, reducing the solution network construction problem to selecting which boundary points to include. For the part of the boundary on the same link end as the solution point, the only possibly relevant path originates from the beginning of the time step, namely $(x_P, t) \to (x_P, t_P) = (x_P, t + \Delta t_{x_P})$. Its constraint is $N(x_P, t_P) \le N(x_P, t) + q_C \Delta t_{x_P}$.

Next, let $L = x_L - x_0$ denote the link length, and let us consider the part of the boundary at the opposite link end. First of all, space-time paths originating from any point B for which

$$t_B > t_P - \begin{cases} \dfrac{L}{\max Z} & \text{if } x_P = x_L \\ \dfrac{-L}{\min Z'} & \text{if } x_P = x_0 \end{cases} \tag{9}$$

should not be included because, by Daganzo (2005a), these paths are not valid as the necessary slope exceeds the range of possible wave speeds $[\min Z', \max Z]$. For the other boundary points, the path to the solution point has a slope of $(x_P - x_B)/(t_P - t_B)$, and, using Daganzo (2005a), its constraint imposed on $N(x_P, t_P)$ thus equals

$$
\begin{aligned}
N(x_P,t_P) &\le N(x_B,t_B)+\int_{t_B}^{t_P}\max_{k\in[0,k_J]}\left(Q(k)-k\frac{x_P-x_B}{t_P-t_B}\right)dt \\
&= N(x_B,t_B)+(t_P-t_B)\max_{k\in[0,k_J]}\left(Q(k)-k\frac{x_P-x_B}{t_P-t_B}\right) \\
&= N(x_B,t_B)+\begin{cases}(x_P-x_B)\max\limits_{k\in[0,k_C]}\left(Q(k)\dfrac{t_P-t_B}{x_P-x_B}-k\right) & \text{if } x_P>x_B \\ (x_P-x_B)\min\limits_{k\in[k_C,k_J]}\left(Q(k)\dfrac{t_P-t_B}{x_P-x_B}-k\right) & \text{if } x_P<x_B\end{cases} \\
&= N(x_B,t_B)+\begin{cases}(x_P-x_B)\max\limits_{q\in\mathrm{dom}\,K}\left(q\dfrac{t_P-t_B}{x_P-x_B}-K(q)\right) & \text{if } x_P>x_B \\ (x_P-x_B)\min\limits_{q\in\mathrm{dom}\,K'}\left(q\dfrac{t_P-t_B}{x_P-x_B}-K'(q)\right) & \text{if } x_P<x_B\end{cases} \\
&= N(x_B,t_B)-\begin{cases}(x_P-x_B)\kappa\left(\dfrac{x_P-x_B}{t_P-t_B}\right) & \text{if } x_P>x_B \\ (x_P-x_B)\kappa'\left(\dfrac{x_P-x_B}{t_P-t_B}\right) & \text{if } x_P<x_B\end{cases} \\
&= \begin{cases}N(x_B,t_B)-\kappa\left(\dfrac{L}{t+\Delta t_{x_L}-t_B}\right)L & \text{if } x_P=x_L \\ N(x_B,t_B)+\kappa'\left(\dfrac{-L}{t+\Delta t_{x_0}-t_B}\right)L & \text{if } x_P=x_0\end{cases}
\end{aligned}
\qquad (10)
$$

Due to Theorem 1 below, we only need to include paths from the opposite boundary with wave speeds within Z (sending flow) or Z' (receiving flow). This is illustrated in Fig. 3.

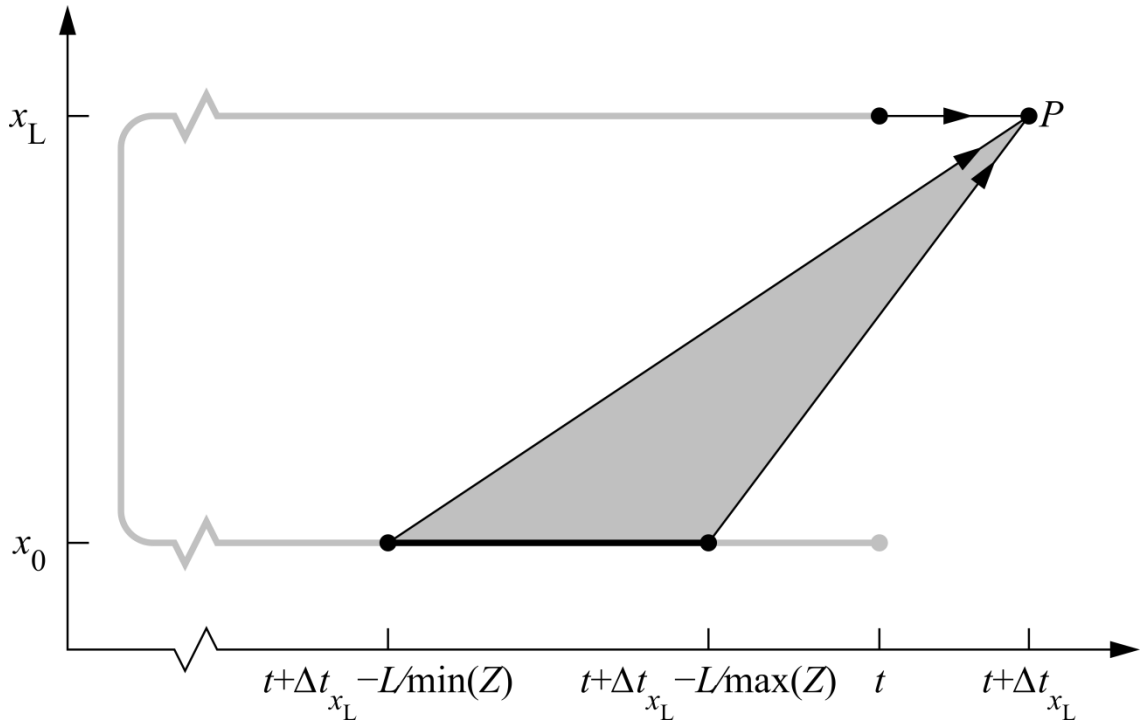

Fig. 3. Solution network for the sending flow, highlighting the relevant part of the boundary condition and the space-time paths corresponding to these origins.

Theorem 1. Paths originating from points (x_B,t_B) on the opposite part of the boundary with $t_B < t_P - L/\min Z$ (sending flow) or $t_B < t_P + L/\max Z'$ (receiving flow) can be excluded from the solution network.

Proof. Consider all points (x_B,t_B) with $t_B \le t_P - L/\min Z$ (sending flow) or $t_B \le t_P + L/\max Z'$ (receiving flow), imposing the constraints

$$
\begin{aligned}
N(x_P,t_P) &\le \begin{cases} N(x_B,t_B) - \kappa\left(\dfrac{L}{t_P-t_B}\right)L & \text{if } x_P = x_L \\ N(x_B,t_B) + \kappa'\left(\dfrac{-L}{t_P-t_B}\right)L & \text{if } x_P = x_0 \end{cases} \\
&= \begin{cases} N(x_B,t_B) - \left(k_C - q_C\dfrac{t_P-t_B}{L}\right)L & \text{if } x_P = x_L \\ N(x_B,t_B) + \left(k_C - q_C\dfrac{t_P-t_B}{-L}\right)L & \text{if } x_P = x_0 \end{cases} \\
&= \begin{cases} N(x_B,t_B) - k_C L + q_C(t_P-t_B) & \text{if } x_P = x_L \\ N(x_B,t_B) + k_C L + q_C(t_P-t_B) & \text{if } x_P = x_0 \end{cases}
\end{aligned}
\qquad (11)
$$

Because $\partial N(x_B,t)/\partial t \le q_C$, the point with $t_B = t_P - L/\min Z$ (sending flow) or $t_B = t_P + L/\max Z'$ (receiving flow) is always the most constraining one. ■

Finally, we observe that, due to the time discretization of the LTM, N is piecewise-linear along the boundary. Therefore we can split the relevant part of the boundary at the opposite link end into a finite set of open intervals with defined, constant flows $\partial N / \partial t$ and a finite set of points. This is illustrated in Fig. 4.

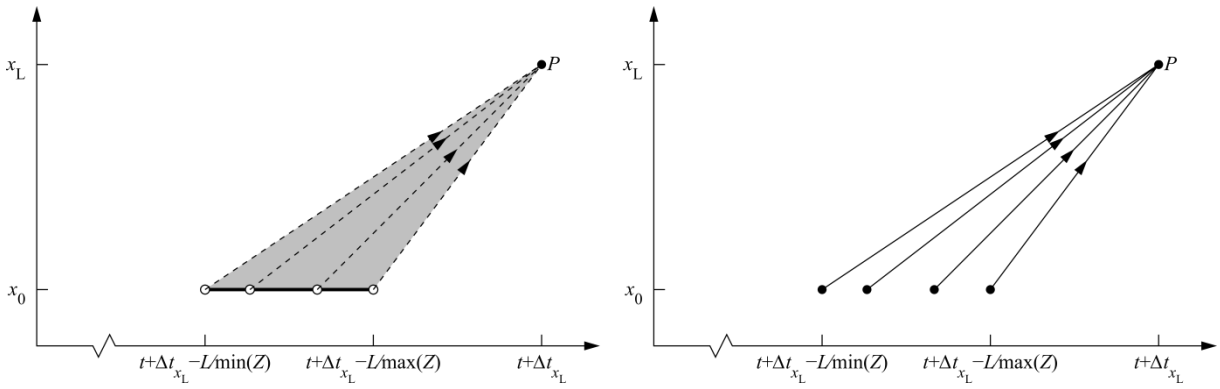

Fig. 4. Opposite boundary of solution network split into open intervals and points.

For the open intervals, instead of investigating all valid paths originating from it, we need only to look at those valid paths with a wave speed compatible with the known boundary flow. More specifically, this means that the wave speed corresponding to the slope from the boundary point to the solution point can be used to construct a tangent line to the FD passing through the traffic state corresponding to the known flow at the boundary. The reason for this is that in order to be constraining, a valid path must be a wave emanated from the boundary (Daganzo, 2005a), and in order for that wave to be emanated, its traffic state must match the traffic state at the boundary. Theorem 2 below shows that this results in only one extra closed-form constraint per open interval, that, together with the above formulas and considerations, allow us to compute $N(x_P,t_P)$ exactly if the boundary conditions are exact.

Theorem 2. If, along the opposite boundary x_B, an open interval (t_1,t_2) with flow $q=(N(x_B,t_2)-N(x_B,t_1))/(t_2-t_1)$ constrains $N(x_P,t_P)$ unlike its infimum t_1 and its supremum t_2, then

$$\begin{cases} t_1+\frac{L}{\min V(q)} < t_P < t_2+\frac{L}{\max V(q)} & \text{if } x_P = x_L \\ t_1-\frac{L}{\max V'(q)} < t_P < t_2-\frac{L}{\min V'(q)} & \text{if } x_P = x_0 \end{cases} \tag{12}$$

and

$$N(x_P,t_P)=\begin{cases} N(x_0,t_1)+q\cdot(t_P-t_1)-K(q)L & \text{if } x_P = x_L \\ N(x_L,t_1)+q\cdot(t_P-t_1)+K'(q)L & \text{if } x_P = x_0 \end{cases}. \tag{13}$$

Proof. The constraint imposed by the interval equals

$$N(x_P,t_P) \le \begin{cases} \inf\limits_{t_B\in(t_1,t_2)} \left(N(x_0,t_1) + q\cdot(t_B-t_1) - \kappa\left(\frac{L}{t_P-t_B}\right)L \right) & \text{if } x_P = x_L \\ \inf\limits_{t_B\in(t_1,t_2)} \left(N(x_L,t_1) + q\cdot(t_B-t_1) + \kappa'\left(\frac{-L}{t_P-t_B}\right)L \right) & \text{if } x_P = x_0 \end{cases}$$

$$\underset{\text{if constraining} \le \text{otherwise}}{=} \begin{cases} \inf\limits_{t_B\in(t_1,t_2)\left|\frac{L}{t_P-t_B}\in V(q)\right.} \left(N(x_0,t_1) + q\cdot(t_B-t_1) - \left(K(q) - q\frac{t_P-t_B}{L}\right)L \right) & \text{if } x_P = x_L \\ \inf\limits_{t_B\in(t_1,t_2)\left|\frac{-L}{t_P-t_B}\in V'(q)\right.} \left(N(x_L,t_1) + q\cdot(t_B-t_1) + \left(K'(q) - q\frac{t_P-t_B}{-L}\right)L \right) & \text{if } x_P = x_0 \end{cases}$$

$$= \begin{cases} \inf\limits_{t_B\in(t_1,t_2)\left|\frac{L}{t_P-t_B}\in V(q)\right.} \left(N(x_0,t_1) + q\cdot(t_P-t_1) - K(q)L \right) & \text{if } x_P = x_L \\ \inf\limits_{t_B\in(t_1,t_2)\left|\frac{-L}{t_P-t_B}\in V'(q)\right.} \left(N(x_L,t_1) + q\cdot(t_P-t_1) + K'(q)L \right) & \text{if } x_P = x_0 \end{cases} \quad . \qquad (14)$$

Since the argument of the infimum is independent of t_B, this reduces to the constraint

$$N(x_P,t_P) \le \begin{cases} \begin{cases} N(x_0,t_1) + q\cdot(t_P-t_1) - K(q)L & \text{if } \exists t_B \in (t_1,t_2) : \frac{L}{t_P-t_B} \in V(q) \\ \infty & \text{otherwise} \end{cases} & \text{if } x_P = x_L \\ \begin{cases} N(x_L,t_1) + q\cdot(t_P-t_1) + K'(q)L & \text{if } \exists t_B \in (t_1,t_2) : \frac{-L}{t_P-t_B} \in V'(q) \\ \infty & \text{otherwise} \end{cases} & \text{if } x_P = x_0 \end{cases} \quad . \qquad (15)$$

Instead of looking whether there exists a point (x_B,t_B) in the interval that can reach (x_P,t_P), we can look whether (x_P,t_P) lies within the total area reached by the interval. This turns the constraint into

$$N(x_P,t_P) \le \begin{cases} \begin{cases} N(x_0,t_1) + q\cdot(t_P-t_1) - K(q)L & \text{if } t_1 + \frac{L}{\max V(q)} < t_P < t_2 + \frac{L}{\min V(q)} \\ \infty & \text{otherwise} \end{cases} & \text{if } x_P = x_L \\ \begin{cases} N(x_L,t_1) + q\cdot(t_P-t_1) + K'(q)L & \text{if } t_1 - \frac{L}{\min V'(q)} < t_P < t_2 - \frac{L}{\max V'(q)} \\ \infty & \text{otherwise} \end{cases} & \text{if } x_P = x_0 \end{cases} \quad . \qquad (16)$$

Finally, if $L/\max V(q) < t_P - t_1 \le L/\min V(q)$ (sending flow) or $-L/\min V'(q) < t_P - t_1 \le -L/\max V'(q)$ (receiving flow), then this constraint of the open interval (t_1,t_2) is equal to the constraint of its infimum t_1. Likewise, if $L/\max V(q) \le t_P - t_2 < L/\min V(q)$ (sending flow) or $-L/\min V'(q) \le t_P - t_2 < -L/\max V'(q)$ (receiving flow), the constraint is equal to that of its supremum t_2. This completes the proof. ■

3.3. Computing sending and receiving flows: algorithms

The following sending and receiving flow algorithms implement the described solution networks. For brevity of notation, we use a compound assignment operator $a :\leq b$ meaning $a := \min(a,b)$ and a floor-to-multiple-of operator $\lfloor a \rfloor_b$ meaning $\lfloor a/b \rfloor \cdot b$.

Algorithm 2. Sending flow.

- *Apply the outflow capacity constraint.*
  $N(x_L, t+\Delta t_{x_L}) := N(x_L, t) + q_C \Delta t_{x_L}$.
- $t_1 := t + \Delta t_{x_L} - L / \min Z$.
- *Apply the constraint of the point* t_1.
  $N(x_L, t+\Delta t_{x_L}) :\leq N(x_0, t_1) - \kappa(\min Z) L$.
- $t_2 := \lfloor t_1 \rfloor_{\Delta t_{x_0}}$.
- Loop:
  - *Find the next* t_2.
    $t_2 := \min(t_2 + \Delta t_{x_0}, t + \Delta t_{x_L} - L / \max Z)$.
  - If $t_1 = t_2$:
    - Exit the loop.
  - $q := \frac{N(x_0, t_2) - N(x_0, t_1)}{t_2 - t_1}$.
  - If $t_1 + \frac{L}{\min V(q)} < t + \Delta t_{x_L} < t_2 + \frac{L}{\max V(q)}$:
    - *Apply the constraint of the open interval* (t_1, t_2).
      $N(x_L, t+\Delta t_{x_L}) :\leq N(x_0, t_1) + q \cdot (t + \Delta t_{x_L} - t_1) - K(q) L$.
  - *Apply the constraint of the point* t_2.
    $N(x_L, t+\Delta t_{x_L}) :\leq N(x_0, t_2) - \kappa\left(\frac{L}{t + \Delta t_{x_L} - t_2}\right) L$.
  - *Move* t_1 *to* t_2.
    $t_1 := t_2$.
- *Set the sending flow*.
  $S(t) := N(x_L, t+\Delta t_{x_L}) - N(x_L, t)$.

Algorithm 3. Receiving flow.

- *Apply the inflow capacity constraint.*
  $N(x_0, t+\Delta t_{x_0}) := N(x_0, t) + q_C \Delta t_{x_0}$.
- $t_1 := t + \Delta t_{x_0} + L / \max Z'$.
- *Apply the constraint of the point* t_1.
  $N(x_0, t+\Delta t_{x_0}) :\leq N(x_L, t_1) + \kappa(\max Z) L$.
- $t_2 := \lfloor t_1 \rfloor_{\Delta t_{x_L}}$.
- Loop:

- *Find the next* t_2 .

  $t_2 := \min\left(t_2 + \Delta t_{x_L}, t + \Delta t_{x_0} + L / \min Z\right)$.

- If $t_1 = t_2$:
  - Exit the loop.
- $q := \frac{N(x_L, t_2) - N(x_L, t_1)}{t_2 - t_1}$.
- If $t_1 - \frac{L}{\max V'(q)} < t + \Delta t_{x_0} < t_2 - \frac{L}{\min V'(q)}$:
  - *Apply the constraint of the open interval* $\left(t_1, t_2\right)$.

    $N\left(x_0, t + \Delta t_{x_0}\right) :\leq N\left(x_L, t_1\right) + q \cdot \left(t + \Delta t_{x_0} - t_1\right) + K'(q) L$.

- *Apply the constraint of the point* t_2 .

  $N\left(x_0, t + \Delta t_{x_0}\right) :\leq N\left(x_L, t_2\right) + \kappa\left(\frac{L}{t + \Delta t_{x_L} - t_2}\right) L$.

- *Move* t_1 *to* t_2 .

  $t_1 := t_2$.

- *Set the receiving flow.*

  $R(t) := N\left(x_0, t + \Delta t_{x_0}\right) - N\left(x_0, t\right)$.

For Algorithm 1, the Courant-Friedrichs-Lewy (1928) condition requires that the time step sizes of nodes are chosen such that for each link,

$$\Delta t_{x_0} \leq \frac{-L}{\min Z'} \wedge \Delta t_{x_L} \leq \frac{L}{\max Z} . \tag{17}$$

Note that these algorithms thus give us exact results according to LWR theory, insofar the input boundary condition is represented exactly as series of time steps with constant flows each.

3.4. Comparison with literature

Now, let us compare this result with discrete-time algorithms previously reported in literature:

- In case of a triangular FD, our algorithms reduce to the algorithms given by Yperman et al. (2006) and Yperman (2007).
- In case of a piecewise-linear concave FD that is not triangular, our algorithms include more constraints on the sending and receiving flows than those proposed by Yperman (2007): we include more paths originating from the boundary at the opposite link end. This difference becomes visible in the model output in case of acceleration fans or rarefaction waves, which the Yperman formulation cannot correctly reproduce. Fig. 5 demonstrates this graphically, where a sudden increase in link inflow results in an outflow spike rather than a monotonically increasing outflow. Bliemer et al. (2016) show a similar example.
- In case of a continuously differentiable concave FD, we can compare our algorithms with the LTM formulation of Gentile (2010). Actually, our algorithms require a discontinuity in the derivative at the capacity point. While this might appear to be a difference with the Gentile formulation, we see upon closer inspection of his algorithms that he constrains the travel times of waves on a link by limiting his look-ahead window. Our restriction on the shape of

the FD makes this implicit constraint explicit. The models are hence similar, e.g. on p. 161-162, Gentile proofs that the constraints resulting from our Theorem 2 represent valid paths (without using that terminology). A difference however is that Gentile uses linear interpolation of cumulative curve in acceleration fans producing constant flows at the opposite link end during the fan. Because each branch of the FD has more than two slopes, this is not correct, as noted by Gentile in his original work. Instead, our algorithms interpolate only within time steps rather than over the entire fan, yielding a more accurate representation of acceleration fans with piecewise-constant flows. A final difference is that our algorithms do not require all nodes to have the same time step size.

- For any concave FD, Mazaré et al. (2011) have previously studied the boundary value problem of LWR theory with N piecewise-linear along the boundary. The main difference with our model is that we do not know the traffic states at the link ends a-priori but compute these endogenously for a network, according to Algorithm 1. Another difference is that we added the previously mentioned restriction to the shape of the FD near the capacity point, enabling us, via Theorem 1, to prevent the computation time from increasing more than linearly with the time horizon of the simulation. Note that our algorithms, derived from variational theory (Daganzo, 2005a) with a solution network with an infinite number of direct paths from the boundary to the solution point, qualify as Lax-Hopf algorithms rather than dynamic programming in the terminology of Mazaré et al. Hence the proposed formulas are both grid-free within the link. The result of our Theorem 2 matches with their Eqs. (24) and (27).

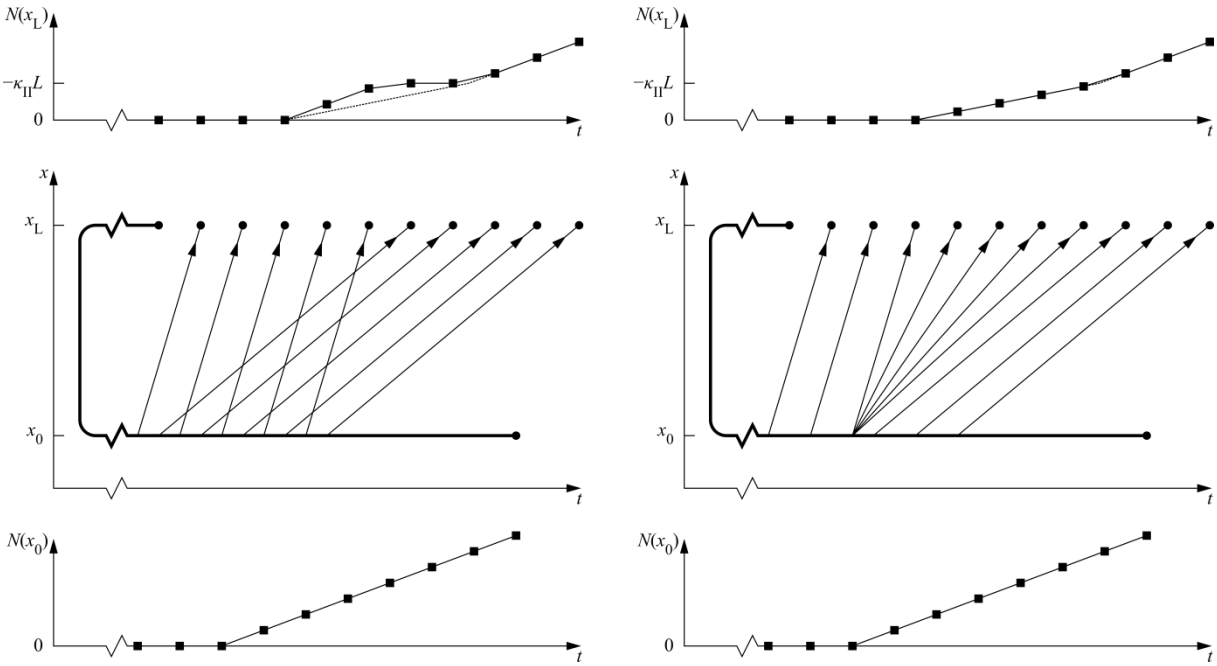

Fig. 5. Reproduction of acceleration fans by the Yperman (2007) model formulation on the left and our model formulation on the right, when both are used with the same piecewise-linear FD. The middle figures indicate the most restrictive paths of each solution network. The analytical solution is indicated with a dashed line in the top figures.

We conclude that according to LWR theory, our proposed algorithms compute the sending and receiving flows exactly on the link level, proven using variational theory. Compared to our

formulation, all variants of the discrete-time LTM previously proposed in literature have either additional restrictions on the shape of the FD, or errors or limitations resulting in inexact solutions, or both.

On the network level, this means that when using our algorithms, the only source of error is the simplification that node flows be constant within time steps, i.e. the time discretization of the boundary conditions that are input to the algorithms above. Since the exactness of the algorithms is naturally subject to the exactness of their input, the network-level solution is generally not exact but converges to the exact solution as the time steps tend to zero, creating a trade-off between numerical accuracy and computational efficiency. For a single invocation of either algorithm, the computational complexity is $O\left(\Delta t^{-1}\right)$. With respect to Δt, this yields a complexity of $O\left(\Delta t^{-2}\right)$ for the dynamic network loading process as a whole: the same as Gentile (2010). These complexities reduce to $O(1)$ and $O\left(\Delta t^{-1}\right)$ respectively in case the FDs are triangular, which are the same as Yperman et al. (2006).

Finally, we remark that in the specific case of a triangular FD, the discrete-time algorithms can be replaced with event-based algorithms as proposed by Raadsen et al. (2016) which are able to produce exact solutions on a network level. However, the current extension of that approach to general concave FDs (Raadsen et al., 2014) incorrectly replaces acceleration fans with shocks, yielding an inadmissible weak solution to the differential equations. Thus, while our discrete-time approach cannot provide the exact network-level solution, it does have the advantage of being able to approximate it by using small time steps. Note that in the future such an approximation may become possible for the event-based approach as well if discretization of FDs into piecewise-linear ones (Bliemer et al., 2016) turns out to be feasible at high resolutions.

4. Capacity drop theory for first-order models

We proceed to extend our model with support for capacity drops. Before we do so, let us review previous work extending first-order models with a capacity drop. As observed by Chung et al. (2007), the capacity drop can be related to the traffic density. Many previously proposed models indeed modify the capacity based on the current density, e.g. in cell transmission models by having the demand function decrease past the critical density (Monamy et al., 2012; Muralidharan et al., 2012; Roncoli et al., 2015). Alvarez-Icaza and Islas (2013) proposed to select the capacity based on the sign of the time derivative of density. In case the capacity can only attain two values, i.e. a free-flow capacity and a queue discharge rate, Srivastava and Geroliminis (2013) proposed to use two threshold densities, where the capacity retains its previous value between the threshold densities and only switches when both thresholds are exceeded. Torné et al. (2014) instead look at whether downstream traffic conditions restricted the flow for a particular location in the previous time step, and set the capacity to the queue discharge rate if so. Jin et al. (2015) specify a kinematic wave theory with capacity drop that constrains the flow over a one-to-one node to the queue discharge rate if congestion is unavoidable.

The previously mentioned models turn out to be difficult to formulate correctly for inhomogeneous roads and general networks. The reason is that the queue discharge rate is defined as the outflow of an active bottleneck, whereas a standing queue is located in front of this bottleneck, on a road with a potentially different free-flow capacity or queue discharge rate itself. For example, if the queue

discharge rate is larger upstream of a bottleneck than downstream, the flow through the bottleneck will be too large if the model is not specified carefully. One could work around this problem by inserting a special transition cell at discontinuities, that could have the free-flow capacity of the upstream cells and the queue discharge rates from the downstream cells. Monamy et al. (2012) indeed propose inserting a special cell to handle the capacity drop at merges. Alternatively, Torné et al. (2014) have cells modify the capacity of neighbors, in addition to using special cells for merges. Such approaches, even if done correctly, are clearly not very useful for extending the LTM, as it does neither discretize links nor nodes into small cells. Jin et al. (2015) are not affected by this issue, but their model formulation can only apply a capacity drop at a pre-specified one-to-one node and neither elsewhere within links nor at more general nodes.

There are however more problems with the driving behavior implied by existing models. For the models that use the average density within each cell to select the cell capacity, the cell size will now influence how the capacity of the road changes over time. This is difficult to understand in terms of driving behavior, which now depends on the discretization of the road, and it is impossible to use these approaches in the LTM due to the lack of such a discretization. Finally, all of the models imply some kind of memory effect in the capacity of a road segment, which may or may not exist in reality. In particular, many models effectively yield a trapezoidal FD when the capacity drop is active, including the extended kinematic wave theory proposed by Jin et al. (2015). As a consequence, traffic at a congested road segment cannot recover to a free-flow state unless the demand for entering that road segment drops and the queue dissolves from its tail. The head of a queue is thus unable to move upstream, while this is observed in reality for so-called wide moving jams or stop-and-go waves. To solve this, one must ensure that the queue discharge state is a point on the free-flow branch. Fortunately, an inverted-lambda style FD (Koshi et al., 1983) can be employed to achieve this, as demonstrated by Hegyi et al. (2008) and Schreiter et al. (2010). However, as pointed out by Torné et al. (2014) and Schreiter et al. (2010), a severe difficulty lies in the possibility of infinite-speed backward shockwaves that can then occur during the onset of congestion. Instead of investigating an inverted-lambda style FD, Lu et al. (2009) mathematically analyzed the solutions of unmodified LWR theory with a simpler jump discontinuity in the FD, which did not result in a capacity drop but did result in the infinite-shockwave-speed problem.

Because of the previous considerations, mere modification of the node model as proposed by Jin et al. (2015) will not be sufficient to describe the capacity drop: the link model must be modified as well, with an inverted-lambda style FD. By extending LWR theory this way, we can correctly account for the capacity drop both when a queue is standing in front of a bottleneck node and when a queue is moving upstream within a link. We will do so in this section.

Firstly, Subsection 4.1 formulates the dynamics for an infinite link in free-flow and in congestion, and Subsection 4.2 defines how the interface between these areas behaves. Next, Subsection 4.3 extends this theory to handle finite links. Subsection 4.4 provides an example of the link dynamics. Next, nodes are discussed in Subsection 4.5. Finally, Subsection 4.6 compares the newly developed traffic flow theory with some of the papers cited above. The results from this section will be added to our model in Sections 5 and 6.

4.1. Link dynamics in free-flow and in congestion

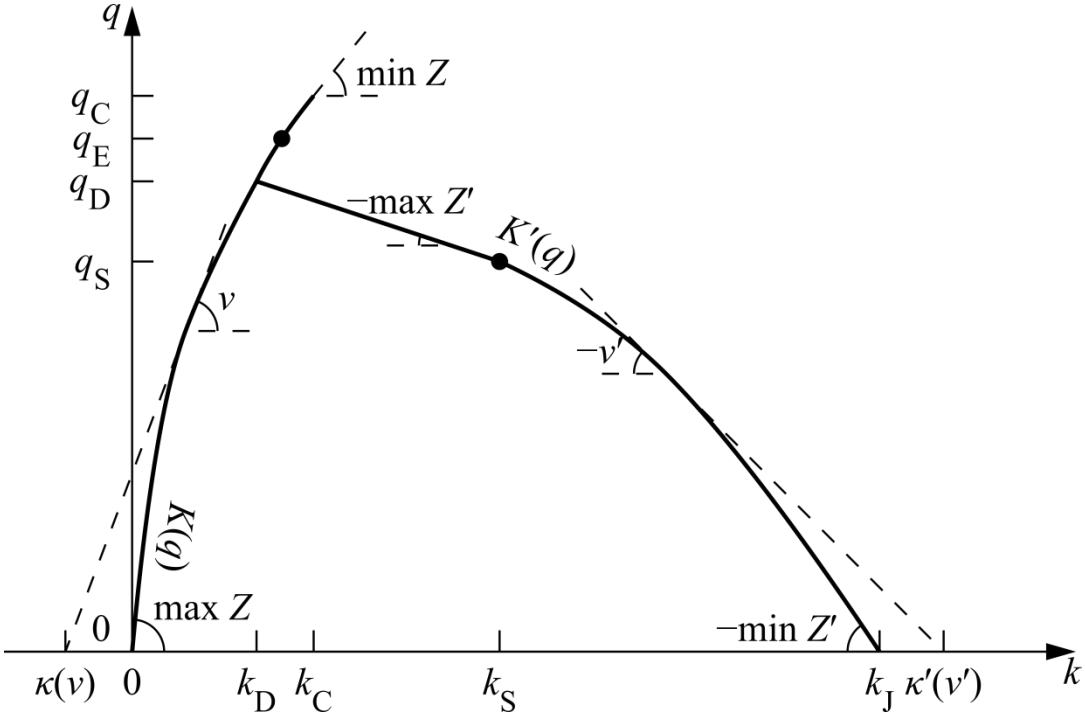

Fig. 6. Example fundamental diagram including a capacity drop. (q_E will be introduced in Subsection 4.5.)

An inverted-lambda style FD is shown in Fig. 6. Clearly, $Q(k)$ ceases to exist as a function. Consequently, we cannot directly use the differential equations from Section 3 to describe traffic flow on a link. However, for space-time areas in free-flow and space-time areas in congestion, we still have the following scalar conservation laws respectively:

$$\begin{aligned} &\frac{\partial q}{\partial x}+\frac{dK(q)}{dq}\frac{\partial q}{\partial t}=0, \quad q\in[0,q_C], \quad K(q)\in[0,k_C] \\ &\frac{\partial q'}{\partial x}+\frac{dK'(q')}{dq'}\frac{\partial q'}{\partial t}=0, \quad q'\in[0,q_D], \quad K'(q')\in[k_D,k_J] \end{aligned} \qquad (18)$$

Note the time and space axes have swapped roles compared to Eq. (6). The equivalent Hamilton-Jacobi equations are

$$\begin{aligned} &\frac{\partial N}{\partial x}+K\left(\frac{\partial N}{\partial t}\right)=0, \quad q=\frac{\partial N}{\partial t}\in[0,q_C], \quad k=-\frac{\partial N}{\partial x}\in[0,k_C] \\ &\frac{\partial N}{\partial(-x)}-K'\left(\frac{\partial N}{\partial t}\right)=0, \quad q'=\frac{\partial N}{\partial t}\in[0,q_D], \quad k'=\frac{\partial N}{\partial(-x)}\in[k_D,k_J] \end{aligned} \qquad (19)$$

Because Hamiltonians $K(\partial N/\partial t)$ and $-K'(\partial N/\partial t)$ are convex in $\partial N/\partial t$, the weak solutions within a free-flow space-time area and within a congested space-time area can be found the same way as before. A definition of the boundary between free-flow and congestion, which provides a boundary condition for both Hamilton-Jacobi equations, now completes our traffic flow theory with capacity drop for the link.

4.2. The interface between free-flow and congestion

To describe the separation between free-flow and congestion, we can use the concept of a separating shock $x_S(t)$, which is a generalized characteristic that divides space-time in a free-flow area upstream and a congested area downstream (Han et al., 2016). Let $N_S(t) = N(x_S(t), t)$ denote the cumulative number of vehicles at the separating shock, and let us describe an infinite link with initial conditions satisfying

$$N(x,0) = N_S(0) + \begin{cases} \int_x^{x_S(0)} K(f(x))dx & \text{for } x < x_S(0) \\ -\int_{x_S(0)}^{x} K'(f(x))dx & \text{for } x > x_S(0) \end{cases} \tag{20}$$

for some $x_S(0)$, $N_S(0)$ and initial flows $f(x)$. This means we have the queue discharge traffic state (k_D, q_D) between congested traffic states and downstream of the last congested state. To ensure that moving observers on both sides of the shock pass vehicles at the same rate, the Rankine-Hugoniot condition (Evans, 2002) requires

$$\begin{aligned} &\frac{dN_S}{dt} = q - k\frac{dx_S}{dt} = q' - k'\frac{dx_S}{dt} \\ &\Rightarrow q - q' = (k - k')\frac{dx_S}{dt} \end{aligned} \tag{21}$$

where (k, q) and (k', q') are the traffic states upstream and downstream of the shock respectively. Let us now state the following differential equations for the separating shock:

$$\begin{aligned} &\lim_{h\downarrow 0} \frac{x_S(t+h) - x_S(t)}{h} = \begin{cases} \frac{q - q'}{k - k'} & \text{if } k' > k_D \\ \max Z & \text{if } k' = k_D \end{cases} \\ &\lim_{h\downarrow 0} \frac{N_S(t+h) - N_S(t)}{h} = q' - k' \lim_{h\downarrow 0} \frac{x_S(t+h) - x_S(t)}{h} \end{aligned} . \tag{22}$$

In the first case, the separating shock is actually a shock and its speed follows from the Rankine-Hugoniot condition. In the second case, the traffic state (k_D, q_D) behind the separating shock lies on the intersection of the free-flow branch and congested branch of the FD, and a speed of $\max Z$ is chosen so that the separating shock is too fast to be an actual shock. Traffic state (k_D, q_D) can then be emanated from the boundary into the free-flow space-time area. Any interaction between traffic state (k_D, q_D) and other free-flow states will thus be handled within the free-flow space-time area by its Hamilton-Jacobi equation, rather than by the separating shock itself.

In the above differential equations, the following definitions ensure that the traffic states upstream and downstream of a space-time point are defined at the shock location:

$$\begin{aligned} &k(x,t) = \lim_{h\downarrow 0} \frac{N(x-h,t) - N(x,t)}{h}, \quad q(x,t) = K^{-1}(k(x,t)) \\ &k'(x,t) = \lim_{h\downarrow 0} \frac{N(x,t) - N(x+h,t)}{h}, \quad q'(x,t) = K'^{-1}(k'(x,t)) \end{aligned} . \tag{23}$$

With the semi-derivatives chosen as above, this forms a causal description of the separating shock and hence of the complete link. While this works fine for FDs with $(k_D, q_D) = (k_C, q_C)$, this leads to two problems if the upstream flow q is larger than the queue discharge rate q_D. The first problem is the possibility that the separating shock violates the Lax (1957) shock admissibility condition, as it may travel upstream too fast for kinematic waves corresponding to downstream state (k', q') to keep up. Hence there exists no solution. The second problem is the infinite-speed shockwave problem described earlier if the downstream flow is also relatively high.

We solve these problems by selecting some fixed point (k_S, q_S) on the congested branch of the FD with $k_C < k_S \leq k_J$ and modifying the shock admissibility criterion for backward shocks whenever an upstream state (k_1, q_1) and a downstream state (k_2, q_2) make contact in space-time, to allow and require shocks emanating traffic state (k_S, q_S) from the shock into the downstream space-time area, provided that

1. this would not violate the Rankine-Hugoniot condition (conservation of vehicles), and
2. there is no solution without a backward shock (flow maximization).

To ensure that this always defines a solution everywhere, it is necessary that the congested branch of the FD is linear between (k_D, q_D) and (k_S, q_S), as in Fig. 6. We then have the following implications for our previous coupling of the two Hamilton-Jacobi systems:

- within the free-flow space-time area, no backward shocks are needed and hence there is no change;
- within the congested space-time area, backward shocks are only needed when $k_1 < k_2$, in which case shocks emanating (k_S, q_S) are not possible without violating the Rankine-Hugoniot condition, so there is no change either;
- at the separating shock, backward shocks are only needed when $q_2 < \min(q_D, q_1)$, and a shock emanating (k_S, q_S) will occur if, subject to the Rankine-Hugoniot condition, the shock between (k_1, q_1) and (k_S, q_S) would diverge from the shock or contact discontinuity between (k_S, q_S) and (k_2, q_2), i.e. if $(q_1 - q_S)/(k_1 - k_S) < (q_2 - q_S)/(k_2 - k_S)$ – otherwise there is no change.

Since the modification of the shock admissibility criterion can only affect the separating shock, let us modify the definition of (k', q') in Eq. (23) as follows, making it dependent on (k, q):

$$k'(x,t) = \begin{cases} k_S & \text{if } x = x_S(t) \wedge \lim_{h \downarrow 0} \frac{N(x,t) - N(x+h,t)}{h} > k_D \wedge \frac{q(x,t) - q_S}{k(x,t) - k_S} < \frac{K'^{-1}\left(\lim_{h \downarrow 0} \frac{N(x,t) - N(x+h,t)}{h}\right) - q_S}{\lim_{h \downarrow 0} \frac{N(x,t) - N(x+h,t)}{h} - k_S} \\ \lim_{h \downarrow 0} \frac{N(x,t) - N(x+h,t)}{h} & \text{otherwise} \end{cases} \quad . \qquad (24)$$

$$q'(x,t) = K'^{-1}(k'(x,t))$$

Here, we eliminated the redundant $q_2 < q_1$ check using $(q_2 - q_S)/(k_2 - k_S) \leq (q_D - q_S)/(k_D - k_S)$ and substituted $(k_1, q_1) = (k(x,t), q(x,t))$ and $(k_2, q_2) = \left(\lim_{h \downarrow 0} \frac{N(x,t) - N(x+h,t)}{h}, K'^{-1}\left(\lim_{h \downarrow 0} \frac{N(x,t) - N(x+h,t)}{h}\right)\right)$ into the

previous condition for emanating (k_S, q_S). When the outcome differs from the previous definition, this changes the propagation of the separating shock per Eq. (22), forming a boundary condition for the congested Hamilton-Jacobi system that emanates (k_S, q_S) from there on where needed in accordance with the above. Note that $k(x,t) > k_D$ is necessary but not sufficient for the second inequality inside Eq. (24) to be true.

One may wonder how to choose the values of (k_S, q_S). This traffic state manifests itself when high-density free-flow traffic breaks down due to downstream congestion. Hence, (k_S, q_S) is the traffic state that occurs in a newly created stop-and-go wave in our model, which can be helpful in choosing appropriate values. A larger k_S means stop-and-go waves become shorter and denser. If one desires to derive (k_S, q_S) empirically, one should consider that our model has infinite deceleration at the back and infinite acceleration at the head of a stop-and-go wave, so one would need to fit reality into this schematic representation of the wave. Additionally, one will wish to take into account that the choice of these values also influences the greatest possible backward speed of information $w < 0$ according to

$$w = \min\left(\frac{q_C - q_S}{k_C - k_S}, \frac{q_C}{k_C - k_J}, \min Z'\right). \tag{25}$$

4.3. Imposing boundary conditions

The previous theory for an infinite link can be extended to a finite link with prescribed boundary values $N(x_0,t)$ and $N(x_L,t)$ by constraining x_S to $[x_0, x_L]$ and adding the following definitions of q and k at x_0 and q' and k' at x_L:

$$\begin{aligned}
q(x_0,t) &= \lim_{h\downarrow 0} \tfrac{N(x_0,t+h)-N(x_0,t)}{h} \\
k(x_0,t) &= K\big(q(x_0,t)\big) \\
q'(x_L,t) &= \begin{cases} q_S & \text{if } x_L = x_S(t) \wedge \lim_{h\downarrow 0} \tfrac{N(x_L,t+h)-N(x_L,t)}{h} < q(x_L,t) \wedge \tfrac{q(x_L,t)-q_S}{k(x_L,t)-k_S} < \tfrac{\lim_{h\downarrow 0} \frac{N(x_L,t+h)-N(x_L,t)}{h} - q_S}{K'\left(\lim_{h\downarrow 0} \frac{N(x_L,t+h)-N(x_L,t)}{h}\right) - k_S} \\ \lim_{h\downarrow 0} \tfrac{N(x_L,t+h)-N(x_L,t)}{h} & \text{otherwise} \end{cases} \\
k'(x_L,t) &= \begin{cases} K'\big(q'(x_L,t)\big) & \text{if } q'(x_L,t) \le q_D \\ k_D & \text{if } q'(x_L,t) \ge q_D \end{cases}
\end{aligned} \tag{26}$$

Similar to before, the dynamics of the separating shock are modified. The main difference is that the necessity of backward shocks is now given by $\lim_{h\downarrow 0} \frac{N(x_L,t+h)-N(x_L,t)}{h} < q(x_L,t)$. Furthermore, note that (k', q') may violate the congested branch of the FD when $x_S = x_L$. This does not matter since at such moments in time, the entire link is and remains in free-flow and the Hamilton-Jacobi equation for the congested space-time area is not needed.

4.4. Example

Fig. 7 illustrates our extension of LWR theory with an example initial-boundary value problem whose solution can be constructed by hand. It has multiple possible choices for $x_S(0)$ that satisfy Eq. (20), but they all result in the same solution for $N(x,t)$. The example has three time periods during which the flow upstream of $x_S(t)$ exceeds q_D, i.e. is equal to q_C. The first two times, (k_S, q_S) is emanated from the separating shock according to Eq. (24). The third time, Eq. (22) initially ensures that the $x_S(t)$ moves downstream. It is constrained to x_L once it reaches the downstream link end, until the link outflow reduces to q_D. The latter causes $x_S(t)$ to start moving upstream, emanating (k_S, q_S), according to Eq. (26).

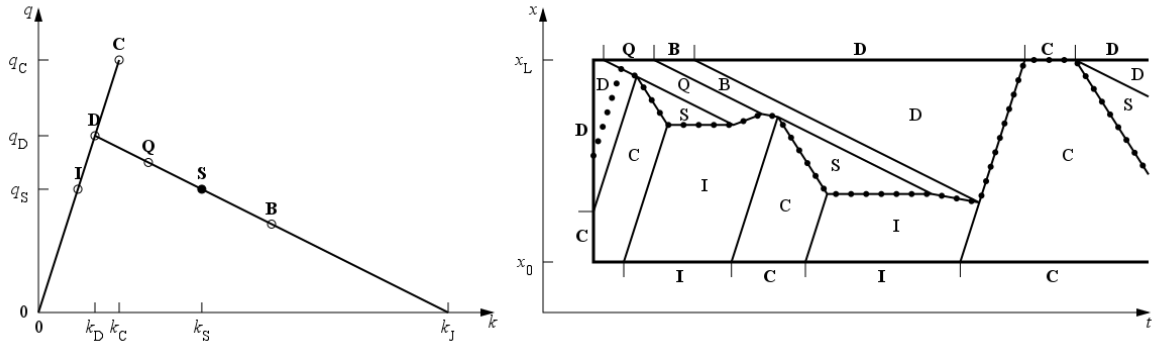

Fig. 7. Example initial-boundary value problem for a link with capacity drop. Letters indicate traffic states in the FD (left) and solution (right). In the right figure, the solution of $x_S(t)$ is dotted for one possible $x_S(0)$.

4.5. Node model requirements

Various first-order node models exist in literature. Our extension of LWR theory with capacity drop does not prescribe a specific one to be used. Nevertheless, the fact that each link may be subject to a capacity drop, adds more constraints to the solution of the node model, so that any chosen node model requires some modifications. These constraints, considering capacity drops for both incoming and outgoing links, are discussed in this subsection.

Firstly, each incoming link cannot transmit a flow between its queue discharge rate and its sending flow, since no appropriate congested traffic state exists to represent such a solution in the inverted-lambda FD of the incoming link. Thus, the transition flow over the node either facilitates the sending flow or is restricted by the queue discharge rate. We will refer to this as the "discharge rate for accelerating" q_D, as it refers to the discharge rate for queues standing or moving on the incoming link, as formulated earlier in this section.

Secondly, it would be illogical if an overloaded outgoing link, i.e. an outgoing link with insufficient receiving flow to accommodate all incoming flows directed at it, would accept more flow than some queue discharge rate, even if permitted by its receiving flow: the overloading implies congestion at

its entrance and congestion implies a restriction to a queue discharge rate, which must be enforced as well. We will refer to this as the "discharge rate for merging" q_E, as it refers to the discharge rate for the entirety of queues standing before the outgoing link with traffic trying to enter it. Similar reasoning may be applied to internal node capacity constraints if one wishes to include those.

Note that these discharge rates q_D and q_E may differ for the same link. This allows to account for different mechanisms behind, and hence different magnitudes of, the capacity drop in both situations. To avoid inconsistencies at redundant one-to-one nodes, we must have $q_D \le q_E \le q_C$ for each link. This is plausible, as there is generally more maneuvering space available at the entrance of a link compared to the interior of the link. Yuan et al. (2016) provide empirical confirmation of $q_D < q_E$ for a motorway link downstream of a lane drop.

Finally, following our considerations in the introduction of this section, we opt to not include memory effects on nodes, thus avoiding that capacity drops in previous time steps affect the maximum node flows in the current time step. Nevertheless, the physical queuing process on an incoming link will always result in the sending flow not exceeding the queue discharge rate for some time if a breakdown occurred in an earlier time step. For consistency of the simulation, ceteris paribus, it is then desirable that the flow in the first time step of the breakdown is the same as in later time steps, also if the queue head is not moving upstream. This desire essentially extends the demand invariance principle of Lebacque and Khoshyaran (2005). Node models with capacity drop should adhere to this principle.

4.6. Comparison with literature

The capacity drop theory proposed in this section has significant advantages over the approach of Lu et al. (2009), who solve the entropy solution for unmodified LWR theory with a discontinuous flow-density relationship. Both feature similar traffic breakdowns; our emanation of traffic state (k_S, q_S) from the separating shock is directly comparable to the deceleration fans in their solution, as derived in Fig. 2 in their paper – it is the same traffic state appearing under the same condition. A difference here is that Lu et al. take the limit $k_S \downarrow k_C$, which creates the infinite-shockwave-speed problem as witnessed by their numerical results, whereas we intentionally chose $k_S > k_C$ to avoid this. Indeed, the breakdown process with $k_S > k_C$ in our model can also be reproduced with unmodified LWR theory with a continuous non-concave FD, as illustrated by e.g. Daganzo (1999, pp. 86-88) who labels (k_S, q_S) a "coasting state".

The most important difference however lies in the discharge process. While our model produces the queue discharge state downstream of congestion, Lu et al. get back the free-flow capacity state there via acceleration shocks, as derived in Fig. 4 and 5 in their paper. Thus, despite their discontinuous FD, their theory still inherits the absence of a capacity drop from unmodified LWR theory, i.e. the maximum flow out of a queue is not less than the maximum flow in free-flow.

We remark that the capacity drop theory proposed by Jin et al. (2015), where queue heads do not move upstream and shocks do not emanate waves, can be interpreted as an extreme case of the more general theory formulated here, with $\max Z' \uparrow 0, q_S \uparrow q_D$ so that, unlike reality, queue heads cannot move upstream and with

$$Q(k)=\begin{cases} K^{-1}(k) & \text{if } k \le k_C \\ \dfrac{k_S-k}{k_S-k_C}q_C+\dfrac{k-k_C}{k_S-k_C}q_S & \text{if } k_C \le k \le k_S \\ K'^{-1}(k) & \text{if } k_S \le k \end{cases} \tag{27}$$

being a concave function, so that traffic state (k_S, q_S) cannot be emanated from a shock.

The node model with capacity drop proposed by Jin et al. (2015) is a special case of the model proposed here, with one incoming link, one outgoing link and a "dropped node capacity" equal to the minimum of the accelerating discharge rate of the incoming link and the merging discharge rates of the outgoing link.

Finally, our capacity drop theory can be seen as a simplified version of the recent proposal by Yuan et al. (2017) who vary the queue discharge rate not only depending on whether the queue is a standing or a moving one, but also on the congestion density, thereby deviating from a fixed inverted-lambda FD. Our simplification is necessary for the link-level solution method that will be developed in the next section. Nevertheless, with a suitable choice of fixed inverted-lambda FDs, our theory exactly reproduces the lane-drop and on-ramp examples of Yuan et al. (2017, p. 481).

5. Link model with capacity drop

In this section, we adapt the solution networks for the receiving and sending flow to implement the link dynamics with capacity drop proposed in the previous section. Rather than computing sending and receiving flows by solving the previous system of differential equations directly, which is cumbersome, we temporarily simplify the problem by not specifying how congestion dissolves. This results in algorithms that are almost as simple as the algorithms presented before in Section 3.3. In Subsection 5.1, we start with investigating the receiving flow; the sending flow is discussed in the Subsection 5.2. These two subsections result in algorithms listed in Subsection 5.3. Finally, consequences and mitigation of the initial simplification will be discussed in Subsection 5.4.

5.1. Computing receiving flows: solution network

The receiving flow $R(t)=N(x_0, t+\Delta t_{x_0})-N(x_0,t)$ will be estimated with a solution network that mimics the above traffic flow theory for an inverted-lambda style FD. Like before, this solution network should consist of two types of paths: paths along the upstream link end enforcing that the link inflow does not exceed the free-flow capacity and paths originating from the downstream link end propagating congested traffic states upstream, but only to the extent such congestion exists at the downstream link end.

To restrict the inflow to the free-flow capacity, one path from (x_0,t) to $(x_0, t+\Delta t_{x_0})$ with cost $q_C \Delta t_{x_0}$, is sufficient. Next, we need to define paths from the downstream link end to $(x_0, t+\Delta t_{x_0})$. We create two types of such paths. Let T denote the first time the downstream link end is congested, which will follow from the output of the node model, to the extent congestion has not dissolved yet within the link. If such a T does not exist, no backward paths need to be considered at all. Otherwise, we firstly create paths corresponding to all wave speeds $v' \in Z'$ for which $t+\Delta t_{x_0} \ge T - L/v'$,

representing all traffic states potentially emanated from the opposite link end. We secondly create additional paths from all $\left(x_L,\hat{t}\right):\hat{t}<T\wedge L/\left(\hat{t}-t-\Delta t_{x_0}\right)\in V'\left(q_S\right)$ as if

$$\forall\hat{t}<T:N\left(x_L,\hat{t}\right)=N\left(x_L,T\right)+q_S\cdot\left(\hat{t}-T\right). \tag{28}$$

This second set of backward paths serves to emanate congested traffic state $\left(k_S,q_S\right)$ from the separating shock if needed, i.e. we simplify the wave emanation from the shock as if these traffic states originated from the downstream link end as well. These paths are collectively enforced by the extra constraint

$$N\left(x_0,t+\Delta t_{x_0}\right)\le N\left(x_L,T\right)+q_S\cdot\left(T-t-\Delta t_{x_0}\right)+k_S L \text{ if } t+\Delta t_{x_0}<T-L\max Z'. \tag{29}$$

Now, Theorem 3 below shows that this solution network is consistent with the Rankine-Hugoniot condition for shocks, allowing us to use the maximum backward information speed w from Eq. (25) to truncate the boundary condition at $t+\Delta t_{x_0}+L/w$ yielding the corresponding time step size constraint $\Delta t_{x_0}\le -L/w$.

Theorem 3. The above solution network solves $N\left(x_0,t+\Delta t_{x_0}\right)$ in a way consistent with shocks satisfying the Rankine-Hugoniot condition.

Proof. This statement is only relevant when T exists, otherwise it is theoretically impossible for $\left(x_0,t+\Delta t_{x_0}\right)$ to be congested, and the solution network contains no backward paths either. If T exists, there will always exist at least one backward path reaching $\left(x_0,t+\Delta t_{x_0}\right)$. The solution network also contains a stationary path along x_0 representing the free-flow regime. Thus, the transition of the upstream link end from free-flow to congestion is implicitly handled by the minimum envelope of both regimes, which is consistent with a Rankine-Hugoniot separating shock. Since shocks within the congested regime are also handled by a minimum envelope, this extends to all shocks. ■

Corollary 1 and Corollary 2 below give us the additional desirable properties that shocks do not emanate $\left(k_S,q_S\right)$ if their inflow does not exceed the queue discharge rate and that receiving flows are non-negative.

Corollary 1. The above solution network does not simulate wave emanation from a shock if the flow into the shock does not exceed the discharge rate q_D.

Proof. Since Theorem 3 ensures the solution network contains no other shocks than Rankine-Hugoniot shocks, a necessary condition for a shock to start emanating $\left(k_S,q_S\right)$ is that its inflow exceeds the queue discharge rate. ■

Corollary 2. The above solution network ensures non-negative receiving flows.

Proof. This statement is trivial if T does not exist. If T exists and is the same as in the previous time step, Theorem 3 proves the statement, since $\partial N/\partial t\ge 0$ in each regime and the Rankine-Hugoniot condition implies continuity of N at the separating shock. The time step size constraint $\Delta t_{x_0}\le -L/w$ ensures that T is known before its existence matters. Corollary

1 ensures that the simulation of wave emanation from the separating shock cannot be constraining immediately after a previous queue has dissolved, so an increase of T cannot cause a negative jump in the permitted N at x_0 either. ■

Finally, Theorem 4 below proves that if there is no capacity drop, this solution network for the receiving flow produces the same result as the solution network derived in Section 3. Hence the solution network proposed here handles both cases with and without a capacity drop.

Theorem 4. The above solution network is equivalent to Algorithm 3 if $(k_D, q_D) = (k_C, q_C)$ regardless of (k_S, q_S).

Proof. Firstly, the Theorem of Daganzo (2005a) permits excluding backward paths corresponding to wave speeds $v' \in [\min Z', 0)$ for which $t + \Delta t_{x_0} < T - L/v'$, without changing the solution. Theorem 1 is not affected by this. Secondly, Corollary 1 ensures that the extra constraint of Eq. (29) is void, removing the dependency on (k_S, q_S). This completes the equivalence. ■

5.2. Computing sending flows: solution network

Our next task is to specify a solution network to estimate the sending flow $S(t) = N(x_L, t + \Delta t_{x_L}) - N(x_L, t)$ mimicking the traffic flow theory for an inverted-lambda style FD. Paths originating from the upstream link end constrain $N(x_L, t + \Delta t_{x_L})$ as in Section 3. However, the constraint enforcing the outflow capacity in case of congestion becomes more complex with a capacity drop: here, the queue discharge rate should be substituted for the capacity, but this constraint should not be active when there is no congestion on the link so that the outflow can attain the free-flow capacity as well.

Let us therefore investigate when the outflow should be constrained to the queue discharge rate. The constraint must be applied after the last time θ the outflow was congested with a flow less than q_D. This constraint will last up till some time $\Theta > \theta$, which will, upon creating congestion in the node model, initially be set to ∞, consistent with our simplification that we ignore the dissolution of congestion. Later, in Subsection 5.4 below, we will replace this with a finite value. The constraint can be implemented as

$$N(x_L, t + \Delta t_{x_L}) \le \begin{cases} N(x_L, \theta) + q_D \cdot (t + \Delta t_{x_L} - \theta) & \text{if } t + \Delta t_{x_L} < \Theta \\ \infty & \text{otherwise} \end{cases} \tag{30}$$

which completes the description of the sending flow algorithm.

Theorem 5, Corollary 3 and Theorem 6 below show that this approach is consistent with Rankine-Hugoniot shocks, produces a sending flow not smaller than zero nor larger than the free-flow capacity and gives the same result as the solution network in Section 3 if there is no capacity drop.

Theorem 5. The solution network described above solves $N(x_L, t + \Delta t_{x_L})$ in a way consistent with shocks satisfying the Rankine-Hugoniot condition.

Proof. Shocks within the free-flow regime and between the free-flow regime and the queue discharge state are handled by a minimum envelope of all valid paths that could be constraining at $\left(x_L, t+\Delta t_{x_L}\right)$. ■

Corollary 3. The solution network described above ensures non-negative sending flows not exceeding the free-flow capacity.

Proof. Each traffic state emitted from the upstream link end, as well as the outflow capacity constraint, has $0 \le \partial N / \partial t \le q_C$. Theorem 5 ensures N is continuous at shocks as well. ■

Theorem 6. The solution network described above is equivalent to Algorithm 2 if $\left(k_D, q_D\right) = \left(k_C, q_C\right)$.

Proof. The equivalence is straightforward for $t+\Delta t_{x_L} < \Theta$, when T exists. For other times, Algorithm 2 has an outflow capacity constraint which the algorithm described above lacks. However, Corollary 3 shows that this absence of the outflow capacity constraint cannot affect the solution. ■

5.3. Computing sending and receiving flows: algorithms

The algorithms listed below implement the link model with support for the capacity drop as previously proposed. As shown above, for $\left(k_D, q_D\right) = \left(k_C, q_C\right)$ they yield solutions identical to the algorithms of the Section 3 without capacity drop. Note that the extended versions presented here use several new variables whose values persist over time steps: Θ, θ and T are persistent real numbers with initial values $\Theta := 0$ and $T := \infty$. Their meanings have been defined in Subsection 5.2, with the addition that we set T to ∞ as a placeholder when it would not exist otherwise. These variables are used by the node model as well, as we see later in Subsection 6.

For brevity of notation, we use a compound assignment operator $a :\le b$ meaning $a := \min\left(a, b\right)$, a floor-to-multiple-of operator $\lfloor a \rfloor_b$ meaning $\lfloor a/b \rfloor \cdot b$ and a ceil-to-multiple-of operator $\lceil a \rceil_b$ meaning $\lceil a/b \rceil \cdot b$.

Algorithm 4. Sending flow with optional capacity drop.

- $t_1 := t + \Delta t_{x_L} - L / \min Z$.
- *Apply the constraint of the point* t_1.
  $N\left(x_L, t+\Delta t_{x_L}\right) :\le N\left(x_0, t_1\right) - \kappa\left(\min Z\right) L$.
- $t_2 := \lfloor t_1 \rfloor_{\Delta t_{x_0}}$.
- Loop:
  - *Find the next* t_2.
    $t_2 := \min\left(t_2 + \Delta t_{x_0}, t + \Delta t_{x_L} - L / \max Z\right)$.
  - If $t_1 = t_2$:
    - Exit the loop.
  - $q := \frac{N\left(x_0, t_2\right) - N\left(x_0, t_1\right)}{t_2 - t_1}$.

- If $t_1 + \frac{L}{\min V(q)} < t + \Delta t_{x_L} < t_2 + \frac{L}{\max V(q)}$:
  - *Apply the constraint of the open interval* (t_1, t_2).
    $N(x_L, t + \Delta t_{x_L}) \coloneqq N(x_0, t_1) + q \cdot (t + \Delta t_{x_L} - t_1) - K(q) L$.
  - *Apply the constraint of the point* t_2.
    $N(x_L, t + \Delta t_{x_L}) \coloneqq N(x_0, t_2) - \kappa\left(\frac{L}{t + \Delta t_{x_L} - t_2}\right) L$.
  - *Move* t_1 *to* t_2.
    $t_1 \coloneqq t_2$.
- If $t + \Delta t_{x_L} < \Theta$:
  - *Apply discharge rate constraint.*
    $N(x_L, t + \Delta t_{x_L}) \coloneqq N(x_L, \theta) + q_D \cdot (t + \Delta t_{x_L} - \theta)$.
- $S(t) \coloneqq N(x_L, t + \Delta t_{x_L}) - N(x_L, t)$.

Algorithm 5. Receiving flow with optional capacity drop[2].

- *Apply the inflow capacity constraint.*
  $N(x_0, t + \Delta t_{x_0}) \coloneqq N(x_0, t) + q_C \Delta t_{x_0}$.
- If $T < t + \Delta t_{x_0} + L / w$:
  - $t_1 \coloneqq \max(T, t + \Delta t_{x_0} + L / \max Z')$.
  - *Emanate congested traffic state* (k_S, q_S) *up to point* t_1 *inclusive.*
    $N(x_0, t + \Delta t_{x_0}) \coloneqq N(x_L, t_1) + q_S (t + \Delta t_{x_0} - t_1) + k_S L$.
  - $t_2 \coloneqq \lfloor t_1 \rfloor_{\Delta t_{x_L}}$.
  - Loop:
    - *Find the next* t_2.
      $t_2 \coloneqq \min(t_2 + \Delta t_{x_L}, t + \Delta t_{x_0} + L / \min Z')$.
    - If $t_1 \geq t_2$:
      - Exit the loop.
    - $q' \coloneqq \frac{N(x_L, t_2) - N(x_L, t_1)}{t_2 - t_1}$.
    - If $t_1 - \frac{L}{\max V'(q')} < t + \Delta t_{x_0} < t_2 - \frac{L}{\min V'(q')}$:
      - *Apply the constraint of the open interval* (t_1, t_2).
        $N(x_0, t + \Delta t_{x_0}) \coloneqq N(x_L, t_1) + q' \cdot (t + \Delta t_{x_0} - t_1) + K'(q') L$.
    - *Apply the constraint for the point* t_2.
      $N(x_0, t + \Delta t_{x_0}) \coloneqq N(x_L, t_2) + \kappa'\left(\frac{L}{t_2 - t - \Delta t_{x_0}}\right) L$.

[2] Use $(k_C, q_C) = (k_D, q_D) = (k_S, q_S)$ and $w = \min Z'$ if there is no capacity drop.

- *Move* t_1 *to* t_2.

  $t_1 := t_2$.

- $R(t) := N\left(x_0, t + \Delta t_{x_0}\right) - N\left(x_0, t\right)$.

For Algorithm 1, the Courant-Friedrichs-Lewy (1928) condition requires that the time step sizes of nodes are chosen such that for each link,

$$\Delta t_{x_0} \leq \frac{-L}{w} \wedge \Delta t_{x_L} \leq \frac{L}{\max Z}. \tag{31}$$

The computational complexity of these algorithms is the same as in Section 3.

5.4. Dissolving congestion

As indicated before, we still need to address the dissolution of congestion. To this end, Fig. 8 shows example output of the previous algorithms with sufficiently small time steps, along with the implied traffic states within the link. Although the onset of congestion correctly follows our traffic flow theory with capacity drop, we can identify three problems related to the resolution of congestion, that occur if the inflow later again exceeds the queue discharge rate. In order of decreasing severity, the problems are that:

1. for large t, the outflow q_D is inconsistent with the inflow q_C, violating our free-flow differential equation;
2. the low inflow q_I has dissolved the queue, but the queue re-appears later, violating our shock admissibility criterion with an incorrectly active separating shock;
3. the low inflow has dissolved traffic state Q within the queue, but this traffic state re-appears later, violating the shock admissibility criterion with an incorrect traffic state behind the separating shock.

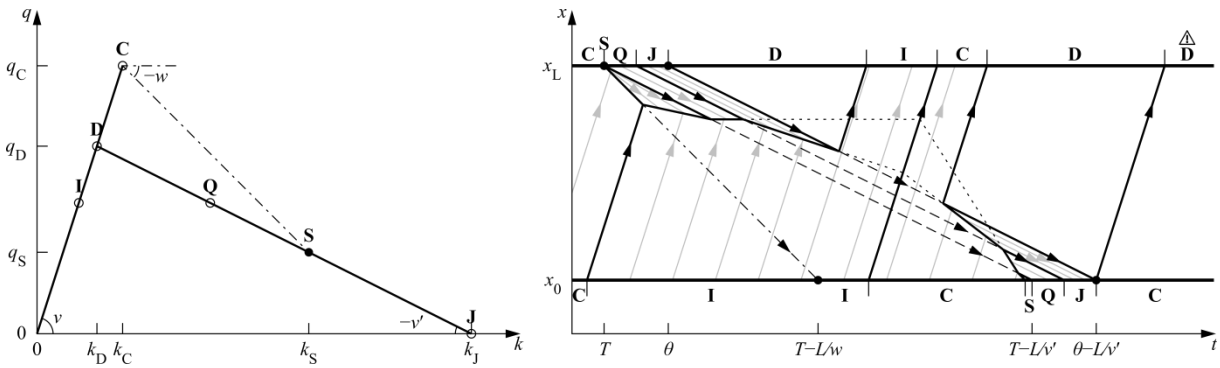

Fig. 8. Problems related to dissolution of congestion. Letters indicate traffic states in the FD (left) and solution (right).

The first problem is solved by setting Θ to its correct value rather than infinity. This also solves the second problem if T is adjusted accordingly. We will investigate how to do so below. It addresses

the third problem as well, but does not solve it entirely if the congested branch of the FD is non-linear.

Essentially, the previous algorithms need to be supplemented with a mechanism that marks congested traffic states as dissolved, recovering the free-flow capacity. Hence, consider a point (x_0,t_0) on the upstream link end and a point (x_L,t_L) on the downstream link end that is preceded by congestion, possibly including the queue discharge state. We will now investigate when we can prove that this congestion will be cleared. Since a queue will always grow if the inflow exceeds the queue discharge rate, assume that the flow prior to (x_0,t_0) does not exceed q_D. Then, in the worst case, the link inflow will be q_C after t_0, such that (x_0,t_0) emanates an acceleration fan. The last traffic state in this fan not exceeding q_D will reach the downstream link end at $t_0+L/\min V(q_D)$, and the corresponding wave will constrain $N(x_L,t_0+L/\min V(q_D))$ to $N(x_0,t_0)+(q_D/\min V(q_D)-k_D)L$. Thus, we know the congestion will be dissolved if

$$N(x_0,t_0)+\left(\frac{q_D}{\min V(q_D)}-k_D\right)L \le N(x_L,t_L)+q_D\cdot\left(t_0+\frac{L}{\min V(q_D)}-t_L\right) \Leftrightarrow$$
$$N(x_0,t_0)-k_DL-q_Dt_0 \le N(x_L,t_L)-q_Dt_L \qquad (32)$$

Since this analysis is based on a valid forward path with speed $\min V(q_D)$, one does not need to verify that the flow prior to (x_0,t_0) is not larger than q_D. Moreover, if the head of a moving jam present before (x_L,t_L) has reached the upstream link end in the time step prior to (x_0,t_0), Eq. (32) becomes an equality so that congestion on the link is also treated as dissolved.

If congestion is dissolved, T can be increased to t_L, ensuring that the dissolved congestion does not constrain the future inflow. If the link is now free of congestion, T can be set to ∞ and Θ can be set to $t_0+L/\min V(q_D)$, ensuring that both later inflow and outflow are unconstrained.

The following algorithm applies this theory, increasing T as congestion is dissolved and incrementing t_0 otherwise, until either of them lies in the future. It must be executed for every link at the beginning of each time step $t:\Delta t_{x_0}\,|\,t \vee \Delta t_{x_L}\,|\,t$, i.e. before computing the sending flow or the receiving flow with the previous algorithms. t_0 is a persistent variable initialized at $-\infty$.

Algorithm 6. Marking congestion as dissolved.

- If $T<t$:
  - Loop:
    - *Ensure that the forward path reaches the downstream link end after T.*
      $t_0 := \max\left(t_0, \left\lfloor T-\frac{L}{\min V(q_D)}\right\rfloor_{\Delta t_{x_0}}+\Delta t_{x_0}\right)$.
    - *Ensure the value of $N(x_0,t_0)$ is already known.*
      If $t_0>t$:
      - Exit the loop.

- *Check whether* (x_0, t_0) *can dissolve congestion at* (x_L, T).

  If $N(x_0, t_0) - k_D L - q_D t_0 \leq N(x_L, T) - q_D T$:

  - *See what more* (x_0, t_0) *could dissolve of this downstream time step.*

    $\hat{T} \coloneqq \min\left(t_0 + \frac{L}{\min V(q_D)}, \lfloor T \rfloor_{\Delta t_{x_L}} + \Delta t_{x_L}\right)$.

  - *Compute the outflow in this downstream time step.*

    $q' \coloneqq \frac{N\left(x_L, \lfloor T \rfloor_{\Delta t_{x_L}} + \Delta t_{x_L}\right) - N\left(x_L, \lfloor T \rfloor_{\Delta t_{x_L}}\right)}{\Delta t_{x_L}}$.

  - *If the outflow is less than* q_D*, not all congestion before* $\left(x_L, \hat{T}\right)$ *necessarily dissolves.*

    If $q' < q_D$:

    - *Thus decrease* $\hat{T}$ *until* $N(x_0, t_0) - k_D L - q_D t_0 \leq N\left(x_L, \hat{T}\right) - q_D \hat{T}$*, by solving the equality.*

      $\hat{T} :\leq \lfloor T \rfloor_{\Delta t_{x_L}} + \frac{N\left(x_L, \lfloor T \rfloor_{\Delta t_{x_L}}\right) - q_D \lfloor T \rfloor_{\Delta t_{x_L}} - N(x_0, t_0) + k_D L + q_D t_0}{q_D - q'}$.

  - *If the downstream time step has been dissolved completely.*

    If $\hat{T} = \lfloor T \rfloor_{\Delta t_{x_L}} + \Delta t_{x_L}$:

    - If $\hat{T} \geq \theta$:

      - *All currently known congestion has been dissolved.*

        $T \coloneqq \infty$.

      - *The outflow should be set to recover to the free-flow capacity once the forward path that solved it reaches the downstream link end.*

        $\Theta \coloneqq t_0 + \frac{L}{\min V(q_D)}$.

      - Exit the loop.

    - Else:

      - *Try dissolving the remainder of this downstream time step with the next forward path.*

        $t_0 \coloneqq t_0 + \Delta t_{x_0}$.

  - *Mark the dissolved congestion as such.*

    $T \coloneqq \hat{T}$.

  - If $T \geq t$:

    - Exit the loop.

- Else:

  - *Try dissolving the congestion with the next forward path instead.*

    $t_0 \coloneqq t_0 + \Delta t_{x_0}$.

6. Node model with capacity drop

As an example of how to adapt a node model to support a capacity drop, we extend the Tampère et al. (2011) node model for unsignalized intersections in this section. The original version of this node model distributes the receiving flows of the outgoing links proportional to the capacities and turning fractions of the incoming links, thereby supporting any number of incoming and outgoing links, maximizing flows to a user-equilibrium and conserving turning fractions. For simplicity, our extended node model will assume that the capacities used for distributing flows are the free-flow capacities.

To accommodate the requirements from Subsection 4.5, the node model is able to modify the sending and receiving flows: a high sending flow is reduced to the accelerating discharge rate if it cannot be transmitted in its entirety, and a high receiving flow is reduced to the merging discharge rate if it is smaller than the sum of the sending flows directed at it. The invariance principle is satisfied, as proven in the Appendix.

The extended algorithm is listed below, following the notation and structure of Tampère et al. (2011). C_i and D_i denote the free-flow capacities and accelerating discharge rates for incoming links integrated over the time step, E_j denotes the merging discharge rate for outgoing links. $\tilde{R}_j$ denotes the unallocated portion of receiving flow R_j, U_j indicates which incoming links are constrained by an outgoing link and J indicates which outgoing links still need to be solved. Calculated in step 2 of the algorithm, C_{ij} denote the oriented capacities, which are the products of the free-flow capacities of incoming links and the corresponding turning fractions. Step 4 finds the most constraining outgoing links $\hat{J} \subseteq J$ [3] and the corresponding reduction factor $\hat{a}$. Then, step 5 first eliminates demand-constrained incoming links ($S_i \leq \hat{a} C_i$). A capacity drop must then occur on the remaining incoming links, so T_i is set to t and their sending flows are capped at their accelerating discharge rates, initializing Θ_i and θ_i. If some sending flows are indeed reduced, step 5 is restarted as there may now exist additional demand-constrained incoming links. Eventually, for the remaining incoming links, the outflows must be smaller than the accelerating discharge rates. If the supply-constrained transition flows G_{ij} together exceed the merging discharge rate E_j of the outgoing link, the entire node model is restarted from step 3, with $R_j = E_j$ to incorporate the capacity drop on the entrance of the outgoing link as well, preserving only previous reductions of sending and receiving flows. If this is not needed, the algorithm simply continues with the next iteration.

Algorithm 7. Node model with optional capacity drops.

1. Retrieve link constraints and initialize supplies and sets.
   - Determine all S_i, S_{ij} and R_j.
   - $\forall i$:
     - $C_i \coloneqq q_{C,i} \Delta t$.

[3] Unlike Tampère et al. (2011), we process multiple outgoing links with identical reduction factors simultaneously instead of sequentially, as the arbitrary order might otherwise influence which capacity drops are activated. Without capacity drops, it would not change any outcome, as identical reduction factors would stay the same until all are processed.

- $D_i \coloneqq q_{D,i}\Delta t$.

- $\forall j$:
  - $E_j \coloneqq q_{E,j}\Delta t$.
  - $\tilde{R}_j \coloneqq R_j$.
  - $U_j \coloneqq \{i \mid S_{ij} > 0\}$.
- $J \coloneqq \{j \mid U_j \neq \varnothing\}$.

2. Determine oriented capacities.

- $\forall i \mid S_i > 0 : \forall j : C_{ij} \coloneqq \frac{S_{ij}}{S_i} C_i$.

3. Stopping criterion.

- If $J = \varnothing$: stop.

4. Determine most restrictive constraints.

- $\forall j \in J : a_j \coloneqq \frac{\tilde{R}_j}{\sum_{i \in U_j} C_{ij}}$.
- $\hat{a} = \min_{j \in J} a_j$.
- $\hat{J} \coloneqq \arg\min_{j \in J} a_j$.

5. Determine flows of corresponding incoming link sets and recalculate receiving flows.

- If $\exists i \in \bigcup_{\hat{j} \in \hat{J}} U_{\hat{j}} \mid S_i \leq \hat{a} C_i$:
  - $\forall i \in \bigcup_{\hat{j} \in \hat{J}} U_{\hat{j}} \mid S_i \leq \hat{a} C_i$:
    - $G_i \coloneqq S_i$.
    - $\forall j : G_{ij} \coloneqq S_{ij}$.
    - $\forall j \in J$:
      - $\tilde{R}_j \coloneqq \tilde{R}_j - G_{ij}$.
      - $U_j \coloneqq U_j \setminus \{i\}$.
      - If $U_j = \varnothing : J \coloneqq J \setminus \{j\}$.
- Else:
  - $\forall i \in \bigcup_{\hat{j} \in \hat{J}} U_{\hat{j}}$:
    - *Mark the time step as congested for the incoming links.*

      $T_i \coloneqq t$.
    - *Activate the queue discharge rate for future time steps.*

      $\Theta_i \coloneqq \infty$.
    - *Set the starting time for applying the accelerating discharge rate.*

      $\theta_i \coloneqq t + \Delta t$.

- *Check whether there are incoming links whose sending flows exceed their accelerating discharge rates, and reduce them to their queue discharge rates as needed.*
  If $\exists i \in \bigcup_{\hat{j} \in \hat{J}} U_{\hat{j}} \mid D_i < S_i$:
  - $\forall i \in \bigcup_{\hat{j} \in \hat{J}} U_{\hat{j}} \mid D_i < S_i$:
    - $\forall j : S_{ij} \coloneqq \frac{S_{ij}}{S_i} D_i$.
    - $S_i \coloneqq D_i$.
  - *Re-check for demand-constrained incoming links since we have reduced sending flows.*
    Restart step 5.
- $\forall i \in \bigcup_{\hat{j} \in \hat{J}} U_{\hat{j}}$:
  - $G_i \coloneqq \hat{a} C_i$.
  - $\forall j : G_{ij} \coloneqq \hat{a} C_{ij}$.
- *Check whether the receiving flow exceeds the merging discharge rate of the outgoing link, and reduce it to its queue discharge rate as needed.*
  If $\exists \hat{j} \in \hat{J} \mid E_{\hat{j}} < \sum_i G_{i\hat{j}}$:
  - $\forall \hat{j} \in \hat{J} \mid E_{\hat{j}} < \sum_i G_{i\hat{j}} : R_{\hat{j}} \coloneqq E_{\hat{j}}$.
  - *Restart the entire algorithm with the reduced receiving flow.*
    $\forall j$:
    - $\tilde{R}_j \coloneqq R_j$.
    - $U_j \coloneqq \{ i \mid S_{ij} > 0 \}$.
  - $J \coloneqq \{ j \mid U_j \neq \varnothing \}$.
  - Go to step 3.
- $\forall i \in \bigcup_{\hat{j} \in \hat{J}} U_{\hat{j}}$:
  - $\forall j \in J$:
    - $\tilde{R}_j \coloneqq \tilde{R}_j - G_{ij}$.
    - If $j \notin \hat{J}$:
      - $U_j \coloneqq U_j \setminus \bigcup_{\hat{j} \in \hat{J}} U_{\hat{j}}$.
      - If $U_j = \varnothing$: $J \coloneqq J \setminus \{ j \}$.
  - $J \coloneqq J \setminus \hat{J}$.
- Go to step 3.

The flow maximization property of the Tampère et al. (2011) model has a special benefit for our extension: it will prevent unnecessary activation of capacity drops. The choice of links for which to activate a capacity drop is not necessarily flow-maximizing however: the order of activation will be

dictated by the reduction factors of the outgoing links, with the most severe reduction first. As an example of non-flow-maximizing behavior, in case of a simple merge, the sending flow of both incoming links is reduced to their accelerating discharge rates simultaneously, while one such reduction might have sufficed.

7. Numerical examples

In this section, we investigate the qualitative properties of the proposed model in numerical examples. We start by demonstrating the elementary improvements of our model over the original LTM with triangular FDs in Subsection 7.1. After this synthetic example, we test the model in a real motorway corridor network in Subsection 7.2.

7.1. Elementary model features

To demonstrate the elementary features of our extended link and node models, we simulate a highway with two on-ramps and visualize the resulting traffic states along the highway. All links have identical FDs. The synthetic demand and FD, with annotated model results for the highway, are shown in Fig. 9, including a comparison with a model with a triangular FD.

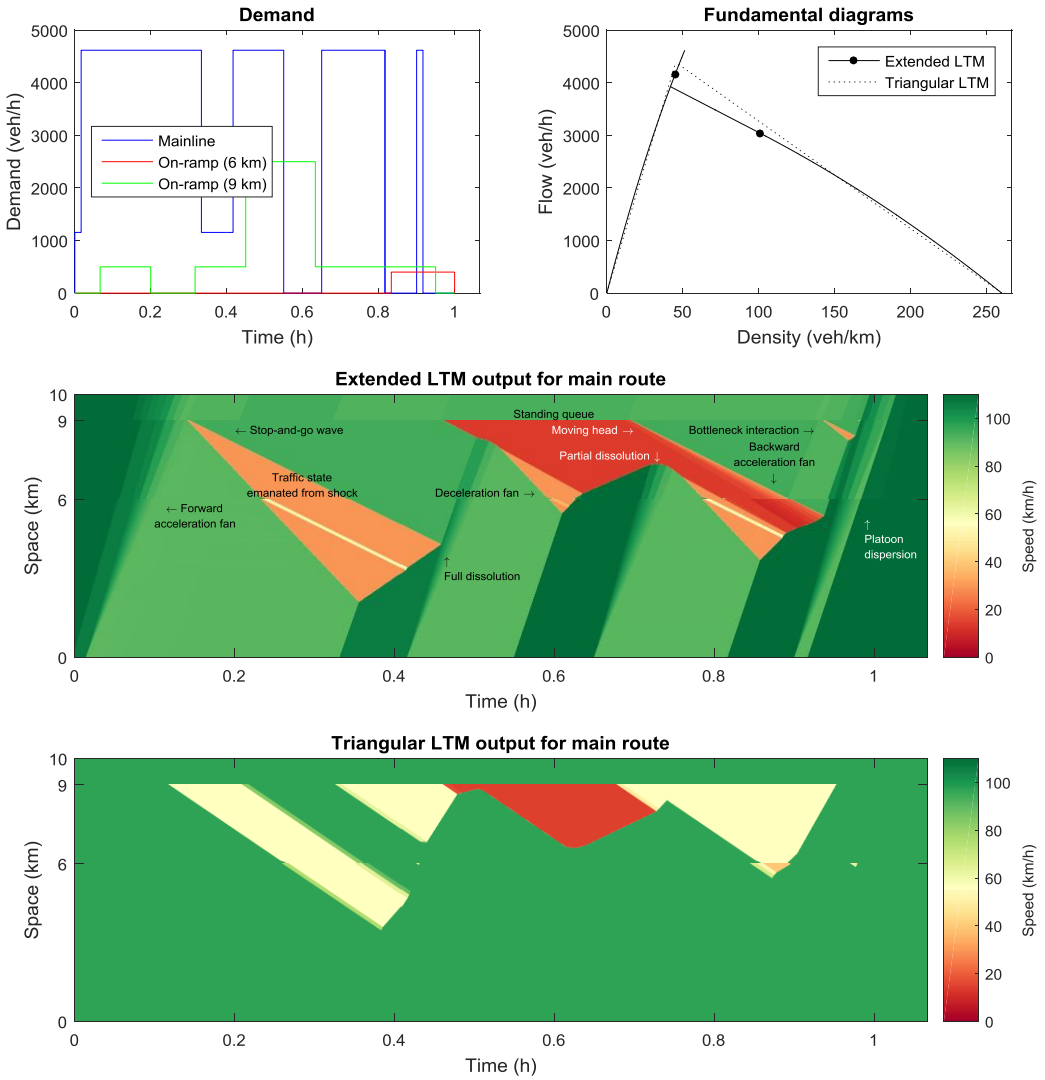

Fig. 9. Simulation input and results for the example highway.

Let us first look at the extended LTM. While the main road's inflow is initially low, it quickly increases to the free-flow capacity, showing a forward acceleration fan. When a low flow from the 9 km on-ramp tries to merge, this generates a stop-and-go wave whose traffic state is emanated from the shock at its tail. The reason that there is no standing queue is that the sum of the accelerating discharge rate from the main road and the low flow from the parallel road do not exceed the free-flow capacity downstream of the merge. Nevertheless, the flow downstream of the merge still drops because the on-ramp flow is not equal to the capacity drop of the stop-and-go wave. A little less than half an hour into the simulation, the stop-and-go wave dissolves due to temporarily low inflow, so that when the inflow is high again later, it can continue without breaking down.

Around the same time, the flow from the ramp increases, causing a standing queue whose outflow equals the merging discharge rate of the link downstream of the merge. Once the queue on the ramp has been dissolved, the head of the queue on the main road starts to move upstream, showing a backward acceleration fan at its head. This is because now the sum of the sending flows does fit into the free-flow capacity downstream, and the node model has no memory of the previous congestion. The queue eventually dissolves, recovering the free-flow capacity of the main road.

Meanwhile, the upstream end of this queue has encountered an inflow higher than its discharge rate, causing it to emanate state (k_s, q_s) from the separating shock. As remarked in Subsection 4.6, this can also be interpreted as a deceleration fan. While this high inflow is temporarily interrupted, the queue partially dissolves, starting a new deceleration fan when the high inflow resumes.

When the head of the queue passes the 6 km on-ramp, which now has a non-zero demand, the flow downstream of this on-ramp increases beyond the accelerating discharge rate of that link: the outflow of the queue remains the same, but the on-ramp can now add more flow. When this higher flow reaches the downstream on-ramp, traffic breaks down again. This means bottlenecks can interact with each other, which can produce an oscillatory traffic pattern even if the demand is constant. We will see more of this in the next subsection. Note that in general the interacting bottlenecks need not be on-ramps. They could include e.g. lane drops, off-ramps, any other geometric discontinuity or even (simplified) capacity funnels (Buckley and Yager, 1974).

Finally, near the end of the simulation, a short platoon with a flow equal to the free-flow capacity originates from upstream. Because the platoon disperses into an acceleration fan, it does not break down when it reaches the 6 km on-ramp even though the on-ramp demand is still non-zero.

If we look at the results of the model with a triangular FD, we see that the acceleration fans are not present and that the speed in free-flow conditions does not decrease with increasing traffic density. Albeit standing queues can dissolve into moving queues if the merging demand drops, such moving queues are unable to grow, since unlike in our extended LTM, the inflow into a moving queue can never exceed its outflow. This also precludes the formation and growth of stop-and-go waves. Finally, the triangular LTM does not reproduce the alternating network outflow pattern of the extended LTM resulting from the capacity drops affecting the merges themselves and the highway upstream of the merges. Note that for both our extended LTM and the triangular LTM, all shocks are crisp, showing that the small numerical error of the original LTM is maintained in our extension.

7.2. Qualitative properties on a motorway corridor network

To investigate the qualitative behavior of the model in a realistic setting, we simulate the Dutch A13 motorway corridor from the Kleinpolderplein interchange near Rotterdam to the Ypenburg interchange near The Hague. This motorway stretch has a length of 12 km, with five off-ramps and six on-ramps. The number of lanes varies between three and four, excluding ramp lanes. We model the corridor as a network of 23 links and 24 nodes as depicted in Fig. 10. We base the capacities on the Dutch motorway capacity manual (Witteveen+Bos and TU Delft, 2011) and construct FDs with a capacity drop of 15% for accelerating and 10% for merging. As a comparison, we also conduct the same simulations with triangular FDs.

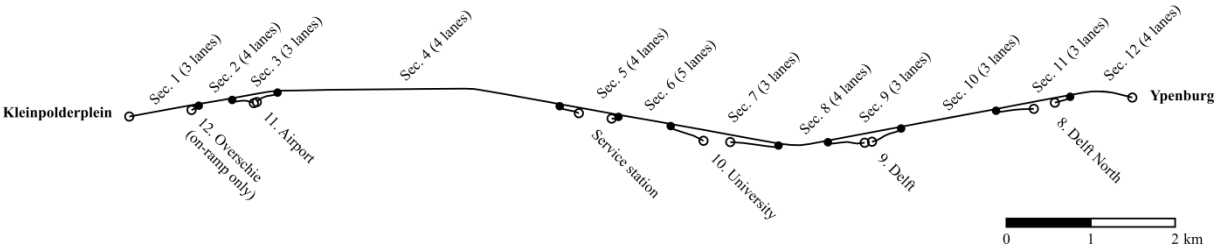

Fig. 10. A13 corridor network.

We simulate three evening peaks in September 2012 with different patterns of congestion. In each simulation, the mainline demand is estimated from one-minute flow measurements of the loop detectors at the beginning of the corridor, and on-ramp demands and off-ramp split fractions are estimated from the differences between flow measurements of loop detectors downstream and upstream of the ramp. This estimation of demand is quite rough and hence both the extended model and the original model show large quantitative errors compared to reality. We nevertheless believe it suffices for the purpose of illustrating the qualitative properties of the model.

The results are plotted in Fig. 11. Qualitatively, we see that our extended LTM replicates the variation of traffic speeds in free-flow conditions, as the used fundamental diagrams contain non-linear free-flow branches. Moreover, the capacity drop in the extended model allows it to reproduce both the onset and propagation of stop-and-go waves, resulting in the structure of the congestion pattern being much closer to reality than the reference model. The extended LTM is able to produce complex mixtures of both moving queues and standing queues, whose composition can be fine-tuned by adjusting the capacity drop.

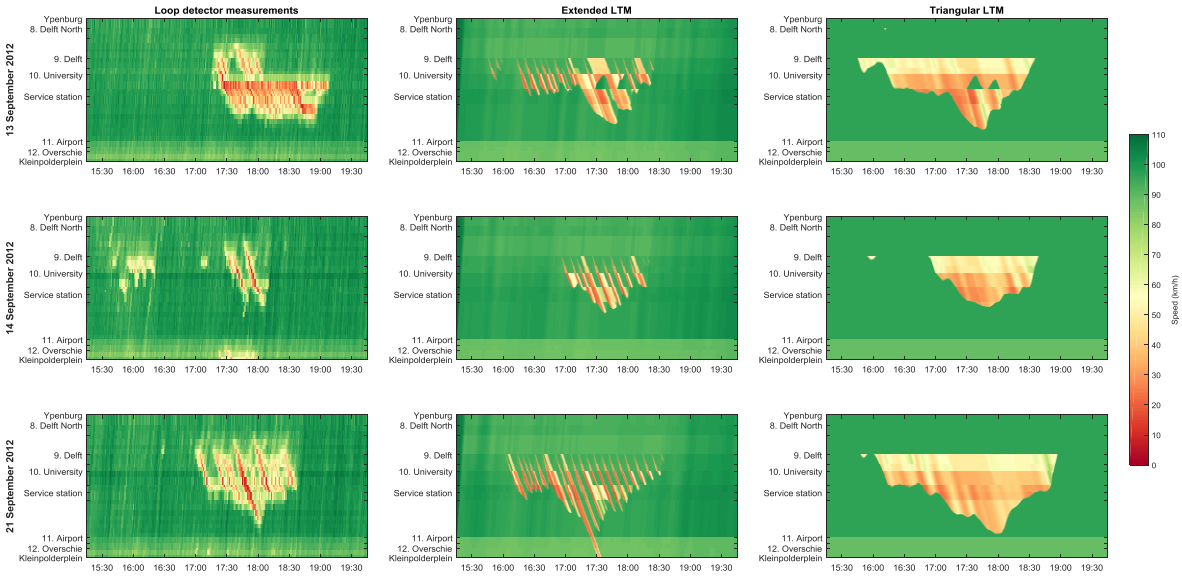

Fig. 11. A13 measurements and simulation results for three evening peaks.

Interestingly, stop-and-go waves generated from an on-ramp can be beneficial for the downstream flow compared to having a standing queue at the same on-ramp, because it results in the on-ramp flow being added on top of the queue discharge rate rather than claiming a part of it. In other words, a moving queue effectively meters the inflow into the original bottleneck. The consequence of this

mechanism is that once the moving queue passes another on-ramp, this metering effect on the original bottleneck will be reduced. This mechanism of interacting bottlenecks, also mentioned in Subsection 7.1, causes repetitive patterns of stop-and-go waves like in reality. The oscillations are a logical consequence of the presence of moving jams and relatively constant demand. The period of oscillation, which is dictated by the distance and wave speeds between the interacting bottlenecks, may be modified by changing the network definition, e.g. by adding explicit capacity funnels.

We also observe that a stop-and-go wave can trigger a standing queue when passing an upstream bottleneck. This again matches reality and it works even though the node model has no memory of capacity drops in previous time steps: the stop-and-go wave creates queues on links in front of the bottleneck, such that after it passed, the accelerating discharge rates of these links exceed the free-flow capacity of the motorway downstream of the bottleneck.

8. Conclusions

In this paper, we extended the LTM, first to handle continuous concave non-triangular FDs, and later to include a capacity drop as well following an extension of LWR theory. For the former case, we showed based on Daganzo (2005a; 2005b) that the link model matches LWR theory up to the accuracy permitted by the discrete time steps with constant link in- and outflows, unlike previously proposed formulations. Next, based on an extension of LWR theory, we modified the solution algorithms of both the link model and the node model to include an optional capacity drop with an inverted-lambda style FD, paying special attention to both the transition from free-flow to congestion and vice versa. Here, the node model is an extension of Tampère et al. (2011), that does not memorize the capacity drops in previous time steps. When the inflow into the node is too high, it first attempts to create stop-and-go waves on the relevant incoming links to reduce the incoming flows to the queue discharge rates, and switches to standing queues, restricted to at most the queue discharge rates of the outgoing links, if the previous reduction would be insufficient. Overall, the result is a computationally-efficient first-order simulation model including optional capacity drops, applicable to general networks, able to produce acceleration fans or rarefaction waves and able to simulate both the onset and propagation of both standing and moving jams.

Both without and with the capacity drop, the numerical error is very small, leading to crisp shockwaves between traffic states. This is an advantage over cell-based models like the CTM and its previously proposed extensions, and also over second-order models with cell-based numerical schemes. Our numerical examples show that our extended model is the first first-order model able to produce traffic patterns featuring the creation, growth and dissolution of both standing queues and stop-and-go waves, consistent with a capacity drop. For further research, we recommend investigating to what extent the qualitative advantages of our extended LTM materialize as quantitative benefits. Since our results show considerable oscillatory interaction between bottlenecks, it would also be relevant to investigate how the model results are affected by details of the network definition, such as capacity funnels, and how this compares to higher-order traffic flow theory.

Acknowledgements

This research effort is funded by the NWO-NSFC project "Optimal Multimodal Network Management for Urban Emergencies", part of the China-Netherlands joint research program "The Application of Operations Research in Urban Transport". The authors wish to thank Yu Han, Kai Yuan and Erik-Sander Smits for the valuable discussions on the topics of this paper. Additionally, we would like to thank the reviewers of this journal for their detailed comments that helped us to improve both the model and the paper significantly.

Appendix

In this appendix, we prove that the node model formulated in Subsection 6 produces the same transition flows in the first time step of a breakdown as in subsequent time steps, ceteris paribus. To this end, we must show that if we reduce a sending flow within the Tampère et al. (2011) algorithm, instead of setting the sending flow directly to this lower value before starting the algorithm, the algorithm produces the same transition flows regardless.

Therefore, let us define a relaxed version of the Tampère et al. (2011) algorithm, such that, at the beginning of step 4a[4], not all incoming links with insufficient demand $\left\{i \in U_{\hat{j}} \mid S_i \le a_{\hat{j}} C_i\right\}$ need to be solved as demand-constrained, but for some reason only a non-empty subset of them (because the final values of S_i might be smaller than the current values). We now show that this change does not affect the outcomes.

Lemma 1. The relaxed algorithm has a finite number of iterations.

Proof. In each iteration, either step 4a removes at least one element from at least one U_j (namely i from $U_{\hat{j}}$), or step 4b removes at least one element from J (namely $\hat{j}$). Since these operations are not reversible and all sets are finite, the stop criterion $J = \varnothing$ is satisfied in a finite number of iterations. ■

Lemma 2. All demand-constrained incoming links are recognized as such by the relaxed algorithm.

Proof. In the relaxed algorithm, step 4a may not empty $U_{\hat{j}}$ and then $\hat{j} \in J \neq \varnothing$ in the next iteration of the algorithm. As proven in Appendix B.1 of Tampère et al. (2011), no a_j will decrease over iterations (this proof also holds for our relaxed algorithm). In later iterations, $a_{\hat{j}} = \min_{j \in J} a_j$ will therefore never be smaller than in previous iterations. This means that all incoming links $i \in U_{\hat{j}}$ satisfying $S_i \le a_{\hat{j}} C_i$ in a previous iteration, but that were not solved as demand-constrained yet, will again satisfy $S_i \le a_j C_i$, regardless of $\hat{j}$ provided that $S_{i\hat{j}} > 0$. Hence, if the algorithm arrives at step 4b, it is guaranteed that $\forall i \in U_{\hat{j}} \mid S_i > a_{\hat{j}} C_i$ and hence no demand-constrained incoming links are still a member of $U_{\hat{j}}$. If step 4b is never reached, $U_{\hat{j}}$ must be empty when the stop criterion is satisfied, which eventually happens due to Lemma 1. Hence all demand-constrained incoming links are removed from all U_j. ■

[4] The equivalent position in our extended algorithm in Subsection 6 is the beginning of step 5. All further references to step numbers in this appendix also refer to Tampère et al. (2011).

Theorem 7. Each transition flow G_i produced by the relaxed algorithm gets the same value as in the original algorithm.

Proof. In each step 3 following step 4, some G_{ij} have been subtracted from the numerator of each a_j and some $C_{ij} \geq G_{ij}$ have been subtracted from the denominator, see Appendix B.1 of Tampère et al. (2011). Since both modifications are commutative, neither the order nor the number of iterations in which step 4 removes incoming links from U_j can impact the first a_j received by step 4b, but only the set of incoming links that are removed. Hence, since Lemma 2 guarantees that at least all demand-constrained incoming links have been removed from U_j, the relaxed algorithm can only fail by removing too many incoming links from U_j. Since, by Tampère et al., a removal cannot decrease a_j, the first a_j in step 4b can only be larger than or equal to the correct value. However, a_j can only get too large after an element from U_j is incorrectly removed, yet the criterion for step 4a ensures that an element can only be removed from U_j incorrectly after a_j is too large. Hence a_j is correct, step 4b produces the correct flows $G_i = a_j C_i$ for $i \in U_j$. If there is a next iteration, the modifications of a_j for all remaining $j \in J$ in step 3 are therefore correct. Because all these modifications commute, the a_j in a next step 4b is again not impacted by the order of removal of incoming links from U_j by either steps 4a or 4b. Again Lemma 2 guarantees that all required removals from U_j have been carried out, and the criterion of step 4a again guarantees that not too many incoming links have been classified as demand-constrained and removed. So a next step 4b also produces correct results. By induction, all steps 4b together produce correct flows for all supply-constrained incoming links. By Lemma 2, the overall algorithm produces correct flows for all demand-constrained incoming links as well. ■

References

Alvarez-Icaza, L., Islas, G.J. (2013). Hysteretic Cell Transmission Model. International IEEE Annual Conference on Intelligent Transportation Systems, 16, pp. 578-583. The Hague.

Bliemer, M.C.J., Raadsen, M.P.H., Bell, M.G.H. (2016). Discrete and continuous time formulations of the link transmission model class. International Symposium on Dynamic Traffic Assignment, 6. Sydney.

Buckley, D.J., Yager, S. (1974). Capacity Funnels near On-Ramps. In: Buckley, D.J. (ed.), Transportation & Traffic Theory. Elsevier, New York, pp. 105-123.

Carey, M., Watling, D. (2012). Dynamic traffic assignment approximating the kinematic wave model: System optimum, marginal costs, externalities and tolls. Transportation Research Part B, 46(5), 634-648.

Chung, K., Rudjanakanoknad, J., Cassidy, M.J. (2007). Relation between traffic density and capacity drop at three freeway bottlenecks. Transportation Research Part B, 41(1), 82-95.

Courant, R., Friedrichs, K., Lewy, H. (1928). Über die partiellen Differenzengleichungen der mathematischen Physik. Mathematische Annalen, 100, 32-74.

Daganzo, C.F. (1994). The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research Part B, 28(4), 269-287.

Daganzo, C.F. (1995). A finite difference approximation of the kinematic wave model of traffic flow. Transportation Research Part B, 29(4), 261-276.

Daganzo, C.F. (1999). The Lagged Cell-Transmission Model. In: Ceder, A. (ed.), Transportation and Traffic Theory; Proceedings of the 14th International Symposium on Transportation and Traffic Theory. Pergamon, Oxford, pp. 81-104.

Daganzo, C.F. (2005a). A variational formulation of kinematic waves: basic theory and complex boundary conditions. Transportation Research Part B, 39, 187-196.

Daganzo, C.F. (2005b). A variational formulation of kinematic waves: Solution methods. Transportation Research Part B, 39, 934-950.

Edie, L.C. (1961). Car-Following and Steady-State Theory for Noncongested Traffic. Operations Research, 9(1), 66-76.

Evans, L.C. (2002). Nonlinear First-Order PDE. In Partial Differential Equations. 19 edn. American Mathematical Society, pp. 94-174.

Gentile, G. (2010). The General Link Transmission Model for Dynamic Network Loading and a Comparison with the DUE algorithm. In: Tampère, C.M.J., Viti, F., Immers, L.H. (eds.), New Developments in Transport Planning: Advances in Dynamic Traffic Assignment. Edward Elgar, pp. 153-178.

Geroliminis, N., Skabardonis, A. (2005). Prediction of Arrival Profiles and Queue Lengths Along Signalized Arterials by Using a Markov Decision Process. Transportation Research Record, 1934, 116-124.

Godunov, S. (1959). A difference scheme for numerical computation of discontinuous solutions of equations of fluid dynamics. Matematischeskii Sbornik, 47(89), 271-306.

Hajiahmadi, M., Corthout, R., Tampère, C., De Schutter, B., Hellendoorn, H. (2013). Variable Speed Limit Control Based on Extended Link Transmission Model. Transportation Research Record, 2390, 11-19.

Hajiahmadi, M., Van de Weg, G.S., Tampère, C.M.J., Corthout, R., Hegyi, A., De Schutter, B., Hellendoorn, H. (2016). Integrated Predictive Control of Freeway Networks Using the Extended Link Transmission Model. IEEE Transactions on Intelligent Transportation Systems, 17(1), 65-78.

Han, K., Piccoli, B., Szetto, W.Y. (2016). Continuous-time link-based kinematic wave model: formulation, solution existence, and well-posedness. Transportmetrica B, 4(3), 187-222.

Hegyi, A., Hoogendoorn, S.P., Schreuder, M., Stoelhorst, H., Viti, F. (2008). SPECIALIST: A dynamic speed limit control algorithm based on shock wave theory. International IEEE Conference on Intelligent Transportation Systems, (pp. 827-832). Beijing.

Himpe, W., Corthout, R., Tampère, C.M.J. (2016). An efficient iterative link transmission model. Transportation Research Part B, 92(B), 170-190.

Jin, W. (2015). Continuous formulations and analytical properties of the link transmission model. Transportation Research Part B, 74, 88-103.

Jin, W., Gan, Q., Lebacque, J. (2015). A kinematic wave theory of capacity drop. Transportation Research Part B, 81(1), 316-329.

Koshi, M., Iwasaki, M., Okhura, I. (1983). Some findings and an overview on vehicular flow characteristics. International Symposium on Transportation and Traffic Flow Theory, 8, pp. 403-426. Toronto.

Lax, P.D. (1957). Hyperbolic Systems of Conservation Laws II. Communications on Pure and Applied Mathematics, 10(4), 537-566.

Lebacque, J., Khoshyaran, M. (2005). First-Order Macroscopic Traffic Flow Models: Intersection Modeling, Network Modeling. In: Mahmassani, H.S. (ed.), Transportation and Traffic Theory: Flow, Dynamics and Human Interaction. Elsevier, pp. 365-386.

Lighthill, M.J., Whitham, G.B. (1955). On kinematic waves II. A theory of traffic flow on long crowded roads. Proceedings of the Royal Society of London A, 229(1178), 317-345.

Long, J., Chen, J., Szeto, W.Y., Shi, Q. (2016). Link-based system optimum dynamic traffic assignment problems with environmental objectives. Transportation Research Part D, in press, doi:10.1016/j.trd.2016.06.003.

Lu, Y., Wong, S.C., Zhang, M., Shu, C. (2009). The Entropy Solutions for the Lighthill-Whitham-Richards Traffic Flow Model with a Discontinuous Flow-Density Relationship. Transportation Science, 43(4), 511-530.

Mazaré, P., Dehwah, A.H., Claudel, C.G., Bayen, A.M. (2011). Analytical and grid-free solutions to the Lighthill-Whitham-Richards traffic flow model. Transportation Research Part B, 45, 1727-1748.

Monamy, T., Haj-Salem, H., Lebacque, J. (2012). A Macroscopic Node Model Related to Capacity Drop. Procedia Social and Behavioral Sciences, 54, 1388-1396.

Muralidharan, A., Horowitz, R., Varaiya, P. (2012). Model Predictive Control of a Freeway Network with Capacity Drops. ASME 2012 5th Annual Dynamic Systems and Control Conference joint with the JSME 2012 11th Motion and Vibration Conference, 2, pp. 303-312. Fort Lauderdale, Florida.

Newell, G.F. (1993). A simplified theory of kinematic waves in highway traffic, part I: General theory. Transportation Research Part B, 27(4), 281-287.

Nie, Y.M. (2011). A cell-based Merchant-Nemhauser model for the system optimum dynamic traffic assignment problem. Transportation Research Part B, 45(2), 329-342.

Papageorgiou, M. (1998). Some remarks on macroscopic traffic flow modelling. Transportation Research Part A, 32(5), 323-329.

Raadsen, M.P.H., Bliemer, M.C.J., Bell, M.G.H. (2014). An efficient event-based algorithm for solving first order dynamic network loading problems. International Symposium on Dynamic Traffic Assignment, 5. Salerno.

Raadsen, M.P.H., Bliemer, M.C.J., Bell, M.G.H. (2016). An efficient and exact event-based algorithm for solving simplified first order dynamic network loading problems in continuous time. Transportation Research Part B, 92(B), 191-210.

Richards, P.I. (1956). Shock Waves on the Highway. Operations Research, 4(1), 42-51.

Roncoli, C., Papageorgiou, M., Papamichail, I. (2015). Traffic flow optimisation in presence of vehicle automation and communication systems - Part I: A first-order multi-lane model for motorway traffic. Transportation Research Part C, 57, 241-259.

Schreiter, T., Smits, E., Van Lint, H., Hoogendoorn, S. (2010). The Cell Transmission Model with Capacity Drop. International TRAIL Congress, 11. Delft.

Srivastava, A., Geroliminis, N. (2013). Empirical observations of capacity drop in freeway merges with ramp control and integration in a first-order model. Transportation Research Part C, 30, 161-177.

Szeto, W.Y. (2008). Enhanced Lagged Cell-Transmission Model for Dynamic Traffic Assignment. Transportation Research Record, 2085, 76-85.

Tampère, C.M.J., Corthout, R., Cattrysse, D., Immers, L.H. (2011). A generic class of first order node models for dynamic macroscopic simulation of traffic flows. Transportation Research Part B, 45(1), 289-309.

Torné, J.M., Soriguera, F., Geroliminis, N. (2014). Coordinated active traffic management freeway strategies using the capacity-lagged cell transmission model. Transportation Research Board Annual Meeting, 93. Washington, D.C.

Van de Weg, G.S., Keyvan-Ekbatani, M., Hegyi, A., Hoogendoorn, S.P. (2016). Urban network throughput optimization via model predictive control using the link transmission model. Transportation Research Board Annual Meeting, 95. Washington, D.C.

Van Driel, C.J.G., Van Arem, B. (2010). The Impact of a Congestion Assitant on Traffic Flow Efficiency and Safety in Congested Traffic Caused by a Lane Drop. Journal of Intelligent Transportation Systems, 14(4), 197-208.

Witteveen+Bos and TU Delft (2011). Capaciteitswaarden Infrastructuur Autosnelwegen. 3rd edn. Dienst Verkeer en Scheepvaart, RW1809-109/houi/012.

Yperman, I.S. (2007). The Link Transmission Model for Dynamic Network Loading. KU Leuven, dissertation.

Yperman, I.S., Logghe, S, Tampère, C.M.J., Immers, L.H. (2006). The Multi-Commodity Link Transmission Model for Dynamic Network Loading. Transportation Research Board Annual Meeting, 85. Washington, D.C.

Yuan, K., Knoop, V.L., Hoogendoorn, S.P. (2017). A kinematic wave model in Lagrangian coordinates incorporating capacity drop: Application to homogeneous road stretches and discontinuities. Physica A, 465, 472-485.

Yuan, K., Knoop, V.L., Leclercq, L., Hoogendoorn, S.P. (2016). Capacity drop: a comparison between stop-and-go wave and standing queue at lane-drop bottleneck. Transportmetrica B, in press, doi:10.1080/21680566.2016.1245163.