Dynamic network loading: a differentiable model that derives link state distributions

Carolina Osorio * Gunnar Flötteröd * Michel Bierlaire *

Report TRANSP-OR 100815
Transport and Mobility Laboratory
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

*Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, {carolina.osoriopizano, gunnar.floetteroed, michel.bierlaire}@epfl.ch

Abstract

We present a dynamic network loading model that yields queue length distributions, accounts for spillbacks, and maintains a differentiable mapping from the dynamic demand on the dynamic queue lengths. The approach builds upon an existing stationary queueing network model that is based on finite capacity queueing theory. The original model is specified in terms of a set of differentiable equations, which in the new model are carried over to a set of equally smooth difference equations. The physical correctness of the new model is experimentally confirmed in several congestion regimes. A comparison with results predicted by the kinematic wave model (KWM) shows that the new model correctly represents the dynamic build-up, spillback, and dissipation of queues. It goes beyond the KWM in that it captures queue lengths and spillbacks probabilistically, which allows for a richer analysis than the deterministic predictions of the KWM. The new model also generates a plausible fundamental diagram, which demonstrates that it captures well the stationary flow/density relationships in both congested and uncongested conditions.

1 Introduction

The dynamic network loading (DNL) problem is to describe the time- and congestion-dependent progression of a given travel demand through a given transportation network. In this article, only passenger vehicle traffic is considered such that the DNL problem becomes to capture the traffic flow dynamics on the road network.

We concentrate on macroscopic models, and we do so for the usual reasons: low number of parameters to be calibrated, good computational performance, mathematical tractability. For reviews on microscopic simulation-based models see, e.g., Hoogendoorn and Bovy (2001), Pandawi and Dia (2005), Brockfeld and Wagner (2006). No matter if macroscopic or microscopic, the modeling of traffic flow has two major facets: the representation of traffic dynamics on a link (homogeneous road segment) and on a node (boundary of several links, intersection).

Models for flow on a link have gone from the fundamental diagram (where flow is a function of density (Greenshields, 1935)) via the Lighthill-Whitham-Richards theory of kinematic waves (where the fundamental diagram is inserted into an equation of continuity (Lighthill and Witham, 1955; Richards, 1956)) to second-order models (where a second equation introduces inertia (Payne, 1971)). Operational solution schemes for both first-order models (e.g., Daganzo, 1995b; Lebacque, 1996) and second-order models (e.g., Hilliges and Weidlich, 1995; Kotsialos et al., 2002) have been proposed in the literature.

Models for flow across a node have been studied less intensively than link models, although they play an important, if not predominant, role in the modeling of network traffic. The demand/supply framework, introduced in Daganzo (1995b) and Lebacque (1996) and further developed in Lebacque (2005), provides a comprehensive foundation for first-order node models. Flow interactions in these models typically result from limited inflow capacities of the downstream links. Recently, this framework has been supplemented with richer features such as conflicts within the node (Flötteröd and Rohde, 2009; Tampere et al., forthcoming).

All models mentioned above are deterministic in that they only capture average network conditions but no distributional information about the traffic states. Arguably, this is so because the kinematic wave model (KWM), the mainstay of traffic flow theory, only applies to average traffic

conditions on long scales in space and time.

Boel and Mihaylova (2006) present a stochastic version of the cell-transmission model (CTM, Daganzo, 1994; Daganzo, 1995a) for freeway segments. Their model contains discontinuous elements, which renders it non-differentiable. Sumalee and co-workers develop a stochastic CTM that approximates traffic state covariances by evaluating a finite mixture of uncongested and congested traffic regimes. The basic model elements can be composed into network structures and are differentiable (Sumalee et al., 2009; Sumalee et al., 2010; Pan et al., 2010). While the CTM constitutes a converging numerical solution procedure for the KWM, it is left unclear to what extent a stochastic CTM converges towards a possibly existing stochastic KWM.

The typical course of action to account for stochasticity in traffic flow models is still to resort to microscopic simulations, which, however, only generate realizations of the underlying distributions and do not provide an analytical framework.

Probabilistic queueing models have been used in transportation mainly to model highway traffic (Garber and Hoel, 2002), and as Hall (2003) highlights: "to this day, problems in highway traffic flow have influenced our understanding of queueing phenomena more than any other mode of transportation." A historical overview of the use of queueing models for transportation is given in Hall (2003). A review of queueing theory models for urban traffic is given in Osorio (2010).

Several simulation models based on queueing theory have been developed, but few studies have explored the potential of the queueing theory framework to develop analytical traffic models. The development of analytical, differentiable, and computationally tractable probabilistic traffic models is of wide interest for traffic management.

The most common approach is the development of analytical stationary models. A review of stationary queueing models for highway traffic is given by Van Woensel and Vandaele (2007). Heidemann and Wegmann (1997) give a literature review for exact analytical stationary queueing models of unsignalized intersections. They emphasize the importance of the pioneer work of Tanner (1962). Heidemann also contributed to the study of signalized intersections (Heidemann, 1994) and presented a unifying approach to both signalized and unsignalized intersections (Heidemann, 1996). Numerical methods to derive the stationary distributions of the main performance measures at an intersection are also given by Oliver and Bisbee (1962), Yeo and Weesakul (1964), Alfa and Neuts (1995) and Viti (2006).

These models combine a queueing theory approach with a realistic description of traffic processes for a given lane at a given intersection. They yield detailed stationary performance measures such as queue length distributions or sojourn time distributions. Nevertheless, they are difficult to generalize to consider multiple lanes, not to mention multiple intersections, or transient regimes. Furthermore, these methods resort to infinite capacity queues and thus fail to account for the occurrence of spillbacks and their effects on upstream links.

Finite capacity queueing theory imposes a finite upper bound on the length of a queue. This allows to account for finite link lengths, which enables the modeling of spillbacks. The methods of Jain and Smith (1997), Van Woensel and Vandaele (2007), and Osorio and Bierlaire (2009c) resort to finite capacity queueing theory and derive stationary performance measures. The models of Jain and Smith (1997) and of Van Woensel and Vandaele (2007) model highway traffic based on the Expansion Method (Kerbache and Smith, 2000), whereas the model of Osorio and Bierlaire (2009c) considers urban traffic and accounts for multiple intersections.

The literature of transient queueing systems is very limited, not to mention the lack of tractable

methods (Peterson et al., 1995), and hence most methods focus on stationary distributions. Accounting for the transients of traffic is necessary to capture in greater detail the build-up and dissipation of spillbacks, and more generally, that of queues. To the best of our knowledge, the work of Heidemann (2001) is the first queueing theory approach to analyze traffic under a transient regime. Stationary performance measures are compared to their transient counterparts, illustrating how they differ and pointing out the importance of accounting for the traffic dynamics. That work also illustrates how the transient performance measures tend with time towards their stationary counterparts. It also indicates that nonstationary models can partially explain the scatter of empirical data, as well as hysteresis loops. Given the complexity of transient analysis, the model of Heidemann (2001) is a classical inifinite capacity queue (M/M/1).

Few methods have gone beyond deriving expected values for the main performance measures, by yielding distributional information (Heidemann, 1994). Distributional information allows to account for the variability of the different performance measures. This is of interest, for instance, when modeling risk-aversion in a route-choice context. Furthermore, as Cetin et al. (2002) mention, classical queueing models do not model the backward wave (also called the negative wave, or jam wave) in congested traffic conditions.

This paper proposes an analytical dynamic queueing model that yields queue length distributions, accounts for spillback, captures the backward wave in congested conditions, and is formulated in terms of boundary conditions that allow for the modeling of dynamic network traffic.

The proposed model builds upon the analytical stationary model of Osorio and Bierlaire (2009b), which resorts to finite capacity queueing theory to describe spillbacks and, more generally, the propagation of congestion. The initial model captures how the queue length distributions of a lane interact with upstream and downstream distributions. Nonetheless, it assumes a stationary regime and thus fails to capture the temporal build-up and dissipation of queues. Here, an analytical transient extension of this model is presented. Adding dynamics to this type of model is a novel undertaking, and we conceive this work to be the first consistent analytical representation of queue length distributions in the DNL problem.

2 Model

Most of this text treats the probabilistic modeling of traffic flow on a homogeneous road segment. Also, the boundary conditions a road segment provides to its up- and downstream node as well as its reaction to the boundary conditions provided by these nodes are developed in detail. Given an additional node model, this enables the embedding of the link model in a general network, which is demonstrated in Section 2.4. In this article, we constrain ourselves to the modeling of nodes with one ingoing and one outgoing link, and we leave the phenomenological modeling of more complex nodes as a topic of future research.

Before presenting the new model, some parallels and differences of the KWM and finite capacity queueing theory are given in Section 2.1. This discussion guides the development of the new model, which consists of a dynamic link model and a static node model. The link model, presented in Section 2.3, is a discrete-time closed-form expression, which guides the transition of the queue distributions from one time step to the next. It is available in closed form under the reasonable assumption of constant link boundary conditions during a time step. No dynamics are introduced into the node model given in Section 2.2, i.e., all node parameters are defined as constant across a

single time step.

2.1 Relation between the KWM and finite capacity queues

As usual, we represent a road by a set of queues, with the main innovation being that the queueing model describes a distribution of the queue length through analytical equations. The comparison of the KWM and finite capacity queueing theory given in this section will serve as a conceptual guideline when developing the new model.

In finite capacity queueing theory, each queue is characterized by:

- an arrival rate, which defines the flow that wants to enter the link from upstream,
- a service rate, which defines the flow that can at most leave the link downstream,
- a queue capacity, which defines how many vehicles fit in the queue.

These parameters have clear counterparts in the demand/supply framework of the KWM (Daganzo, 1994; Lebacque, 1996). The arrival rate corresponds to the flow demand (typically denoted by Δ) at the upstream end of the link. The service rate corresponds to the flow supply (typically denoted by Σ) at the downstream end of the link. Finally, the queue capacity is directly related to the length of the link and its jam density.

These symmetries, however, are imperfect. In particular, consistent solutions of the KWM are known to satisfy the invariance principle (Lebacque, 2005), which essentially states that the flow is not affected by

- increasing the upstream demand in congested conditions or
- increasing the downstream supply in uncongested conditions.

The invariance principle does not hold in finite capacity queueing theory. This is because the flow between two queues is treated as a vehicle transmission event that occurs with a probability that basically results from multiplying (i) the probability that the upstream queue is non-empty and (ii) the probability that the downstream queue is non-full. Changing any of these probabilities also changes their product, and it does so both in free-flow and congested traffic conditions.

On a side note, these considerations suggest that the invariance principle is more generally at odds with possible stochastic versions of the KWM. Consider the basic flow transmission rule $q = \min\{\Delta, \Sigma\}$, where q is the flow and Δ and Σ are the demand and the supply. Evaluating the expectation of q for distributed Δ and Σ, one obtains

$$\mathrm{E}\{q\} = \int\int \min\{\Delta, \Sigma\} p(\Delta, \Sigma) d\Delta d\Sigma \tag{1}$$

where $p(\cdot)$ represents the joint probability density function of demand and supply. Possible inconsistencies with the invariance principle result from the fact that the above expectation can mix free-flow and congested traffic conditions.

2.2 Node model

We model a set of links in series. Vehicles arrive to the first link, travel along all links and leave the network at the last link. To formulate the node model, we introduce the following notation:

i	link index, numbered consecutively from 1 in the direction of flow;
k	time interval index;
$q_i^{\text{in},k}$	inflow to link i during time interval k (in vehicles per time unit);
$q_i^{\text{out},k}$	outflow from link i during time interval k (in vehicles per time unit);
μ_i^k	flow capacity of the downstream node of link i during time interval k (in vehicles per time unit);
N_i^k	number of vehicles in link i during time interval k;
ℓ_i	space capacity (maximum number of vehicles) of link i.

The KWM predicts an expected flow $\min\{\Delta_i^k, \Sigma_{i+1}^k\}$ between two links where Δ_i^k is the expected demand from the upstream link i and Σ_{i+1}^k is the expected supply provided by the downstream link $i+1$, all in time interval k. In finite capacity queueing theory, the flow between two links results from vehicle transmission events that occur with the probability that the upstream queue is non-empty and the downstream queue is non-full, that is with probability

$$P(N_i^k > 0,\ N_{i+1}^k < \ell_{i+1}), \tag{2}$$

where

- $N_i^k > 0$ is the event that there is at least one vehicle in the upstream queue i in time interval k, i.e., there is at least one vehicle ready to leave the upstream link;
- $N_{i+1}^k < \ell_{i+1}$ is the event that the downstream queue is not full in time interval k, i.e., there is no spillback from downstream.

Given the node model parameters, we assume the link models to be independent. Thus,

$$P(N_i^k > 0,\ N_{i+1}^k < \ell_{i+1}) = P(N_i^k > 0)P(N_{i+1}^k < \ell_{i+1}). \tag{3}$$

This is a simplification of real traffic phenomena: splitting a link into half and inserting a node of this type would remove any spatial correlation between the upstream and downstream half of that link. For now, we maintain the independence assumption in order to arrive at a first instance of a full network model, where we minimize its side-effects by limiting the locations of nodes to true intersections and modelings links as uninterrupted entities.

The flow across the node is then given by the product of these probabilities with the node capacity μ_i^k:

$$q_i^{\text{out},k} = \mu_i^k P(N_i^k > 0)P(N_{i+1}^k < \ell_{i+1}). \tag{4}$$

That is, the flow reaches μ_i^k when the link configurations are such that vehicle transitions occur with probability one.

The node capacity μ_i^k captures both the link's flow capacity (resulting from, e.g., its free-flow speed and number of lanes) and intersection attributes (e.g., signal plans, ranking of traffic streams).

In previous work, it has been determined based on national transportation standards such as the Swiss VSS norms or the US Highway Capacity Manual (Osorio and Bierlaire, 2009b).

Flow conservation defines the inflow of a given link as the outflow of its upstream link, i.e.,

$$\forall i > 1,\ q_i^{\text{in},k} = q_{i-1}^{\text{out},k}. \tag{5}$$

The inflow of the exogeneous demand into the first link and the outflow of the last link are described in the more general context of Section 2.4. This formulation can be extended to allow for arbitrary link topologies as well as more general demand structures (where external arrivals and departures arise at arbitrary links). These extensions can be based, for instance, on the assumptions of the Osorio and Bierlaire (2009a) approach.

2.3 Link model

In this section, we describe how we derive the queue length probability distributions. We also describe how a link is represented by a set of queues.

2.3.1 Finite capacity queueing model

We build upon the urban traffic model of Osorio and Bierlaire (2009b). A formulation for large-scale networks appears in Osorio (2010). Both of these models are derived from the analytical stationary queueing model of Osorio and Bierlaire (2009a).

We briefly recall the main components of the stationary queueing model. This analytical model considers an urban road network composed of a set of both signalized and unsignalized intersections. Each link is modeled as a set of queues. The road network is therefore represented as a queueing network. It is analyzed based on a decomposition method, where performance measures for each queue, such as stationary queue length distributions and congestion indicators, are derived.

In order to account for the limited physical space that a queue may occupy, the model resorts to *finite capacity queueing theory*, where there is a finite upper bound on the length of each queue. The use of a finite bound allows to capture the impact of queues on upstream queues (i.e., spillbacks) and to consider scenarios where traffic demand may exceed supply. In queueing theory terms, this corresponds to a traffic intensity that may exceed one. These are the main distinctions between classical queueing theory and finite capacity queueing theory.

The initial model describes the between-queue interactions. Congestion and spillbacks are modeled by what is referred to in queueing theory as *blocking*. This occurs when the queue length reaches its upper bound and thus prevents upstream vehicles from entering the queue, i.e., it blocks arrivals from upstream queues at their current location. This blocking process is described by endogenous variables such as blocking probabilities and unblocking rates. In particular, the probability that a queue spills back corresponds to the *blocking probability* of a queue.

All distributional assumptions and approximations of the Osorio and Bierlaire (2009a) model are preserved in the new framework. A detailed discussion is given in Osorio (2010). In particular, classical assumptions are used to ensure tractability and to allow for closed-form expressions. For a given queue, the inter-arrival times, the service times, and the times between successive unblockings (events of a previously blocked queue becoming available again) are assumed exponentially distributed and independent random variables. These assumptions enable a tractable transient queueing analysis

(Newell, 1979) but leave room for improvements, see also the discussion of Equation (3) in Section 2.2.

2.3.2 Dynamic queueing model

The stationary model derives the queue length distributions from the standard queueing theory *global balance equations*. Coupling equations are used to capture the network-wide interactions between these single-queue models. The new dynamic version of this model consists of a dynamic link model and a static node model. The global balance equations are replaced by a continuous-time closed-form expression for the transient queue length distributions.

This model is implemented in discrete-time, i.e., the dynamic expression guides the link model's transition from the queue length distribution of one time step to the next. It is available in closed form under the (reasonable) assumption of constant link boundary conditions during a simulation step (Morse, 1958). No dynamics are introduced into the node model, which maintains the structure of the original stationary model.

We introduce the following notation:

δ	time step length;
$p_{i,n}^k(t)$	transient probability that queue i is of length n at continuous time t of time interval k, where t is in $[0, \delta]$;
$s_{i,n}^k$	stationary probability that queue i is of length n during time interval k;
ρ_i^k	traffic intensity of queue i during time interval k;
$\hat{\mu}_i^k$	service rate of queue i during time interval k (differs from the node service rate μ_i^k in that it accounts for spillback from downstream, see Equation (10));
λ_i^k	arrival rate of queue i during time interval k;

Each queue is defined based on three parameters: the arrival rate λ_i^k, the service rate $\hat{\mu}_i^k$, and the upper bound on the queue length ℓ_i. The ratio of arrival to service rates (i.e., of demand to supply) is known in queueing theory as the *traffic intensity*, which is given by

$$\rho_i^k = \frac{\lambda_i^k}{\hat{\mu}_i^k}. \tag{6}$$

As described in Section 2.3.1, for finite capacity queues the traffic intensity is unbounded and in particular may exceed one. This allows for highly congested traffic conditions where demand exceeds supply.

Given ℓ_i and ρ_i^k, the stationary queue length distribution is given in closed-form as

$$s_{i,n}^k = \frac{1 - \rho_i^k}{1 - (\rho_i^k)^{\ell_i+1}} (\rho_i^k)^n. \tag{7}$$

This expression holds for the type of queues considered in this framework, which are known as M/M/1/ℓ queues. Details on the derivation of these stationary probabilities appear in Bocharov et al. (2004).

The transient probabilities for a given queue i and continuous time t from 0 to δ within a given time interval k have been derived by Morse (1958), Equation (6.13), to be

$\forall n = 0, 1, ..., \ell_i, \forall t \in [0, \delta]$

$$\begin{cases} p_{i,n}^k(t) = s_{i,n}^k + (\rho_i^k)^{\frac{n}{2}} \sum_{j=1}^{\ell_i} C_{i,j}^k \left\{ \sin\left(\frac{jn\pi}{\ell_i+1}\right) - \sqrt{\rho_i^k} \sin\left(\frac{j(n+1)\pi}{\ell_i+1}\right) \right\} e^{\tau_{i,j}^k t} & \text{(8a)} \\ \tau_{i,s}^k = \lambda_i^k + \hat{\mu}_i^k - 2\sqrt{\lambda_i^k \hat{\mu}_i^k} \cos\left(\frac{s\pi}{\ell_i+1}\right), & \text{(8b)} \end{cases}$$

where the coefficients $\{C_{i,j}^k\}_j$ are chosen to fit the initial values of the transient distribution at the beginning of the time interval by solving an ℓ_i-dimensional linear system of equations such that

$$p_{i,n}^k(0) = p_{i,n}^{k-1}(\delta) \tag{9}$$

holds. Compliance with Equation (9) maintains the temporal continuity of the queue length distributions.

Let us now describe how the arrival and service rates of a queue are associated to the underlying link attributes.

Service rate The service rate of a queue i during time interval k is often referred to as the *effective service rate* (Osorio and Bierlaire, 2009a) and is given by

$$\hat{\mu}_i^k = \mu_i^k P(N_{i+1}^k < \ell_{i+1}). \tag{10}$$

As stated in Section 2.2, μ_i^k accounts for the flow capacity of the underlying link and its downstream node. Additionally, the effective service rate $\hat{\mu}_i^k$ accounts for flow reductions due to spillbacks from downstream. This is captured through the probability $P(N_{i+1}^k < \ell_{i+1})$ that the downstream link does not spill back.

Arrival rate The type of queueing models used in this work are known as *loss models* (see Osorio (2010) for a description of these models). For such models, the inflow to a queue and its arrival rate are related as follows:

$$q_i^{\text{in},k} = \lambda_i^k P(N_i^k < \ell_i), \tag{11}$$

i.e., the inflow to the link corresponds to those arrivals that occur while the link is not full, which occurs with probability $P(N_i^k < \ell_i)$. The notion of a "loss" model should not be taken literally; vehicles that are unable to enter a full link are stored in the upstream link and are not discarded.

2.3.3 Full link model

According to the node model (Equations (4) and (5)), a link provides two boundary values to its adjacent nodes. The first is the probability $P(N_i^k > 0)$ that there are vehicles at the downstream end of the link ready to proceed downstream. The second is the probability $P(N_i^k < \ell_i)$ that the link does not spill back.

In order to calculate these two probabilities, we model each lane of a link as a set of two queues, referred to as the upstream queue (UQ) and the downstream queue (DQ). These queues are depicted in Figure 1 and explained in the following.

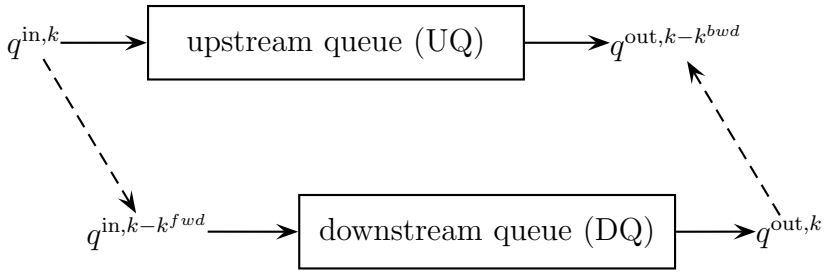

Figure 1: Link modeled with two time-shifted queues

First, the role of the downstream queue is to capture the downstream dynamics of the link, which define $P(N_i^k > 0)$. This downstream queue considers all vehicles in the link that are ready to leave the link, i.e., all vehicles that are in its physical queue.

To accurately represent the length of the DQ, we account for the time needed by a vehicle to traverse the link in free-flow conditions. As is depicted in Figure 1, the inflow to the DQ during time interval k, $q_i^{\text{in DQ},k}$, is equal to the inflow of the link lagged by a fixed number of time intervals, k^{fwd}, which represents the free-flow travel time, i.e.,

$$q_i^{\text{in DQ},k} = q_i^{\text{in},k-k^{fwd}}. \tag{12}$$

This ensures finite vehicle progressions in uncongested conditions. Furthermore, the outflow of the DQ corresponds to the link outflow:

$$q_i^{\text{out DQ},k} = q_i^{\text{out},k}. \tag{13}$$

Second, the role of the upstream queue is to capture the upstream dynamics of the link in order to derive the probability $P(N_i^k < \ell_i)$ that there is no spillback. This queue captures the finite dissipation rate of vehicular queues: upon the departure of a vehicle from a link it accounts for the time needed for this newly available space to reach the upstream end of the link.

This is achieved by setting the outflow $q_i^{\text{out UQ},k}$ of the UQ during time interval k equal to that of the link lagged by a constant k^{bwd}, which represents the time needed for the available space to travel backwards and reach the upstream end of the link:

$$q_i^{\text{out UQ},k} = q_i^{\text{out},k-k^{bwd}}. \tag{14}$$

The inflow of the UQ is equal to the link inflow:

$$q_i^{\text{in UQ},k} = q_i^{\text{in},k}. \tag{15}$$

The upstream queue UQ accounts for all vehicles that are on the link as well as those that have recently left but their corresponding available space has not yet reached the upstream end of the link. This queue is necessary to correctly capture the congested half of the fundamental diagram, which is elaborated further below.

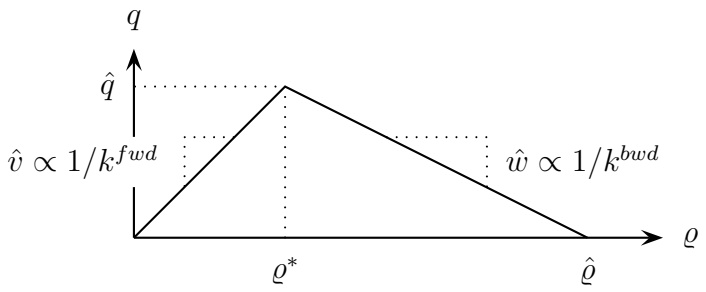

Figure 2: Fundamental diagram for deterministic double-queue model

The boundary conditions $P(N_i^k > 0)$ and $P(N_i^k < \ell_i)$ provided by the link are obtained from the downstream and the upstream queueing models through

$$P(N_i^k > 0) = 1 - p_{i,0}^{\mathrm{DQ},k}(0) = 1 - p_{i,0}^{\mathrm{DQ},k-1}(\delta) \tag{16}$$

and

$$P(N_i^k < \ell_i) = 1 - p_{i,\ell_i}^{\mathrm{UQ},k}(0) = 1 - p_{i,\ell_i}^{\mathrm{UQ},k-1}(\delta) \tag{17}$$

where $p_{i,0}^{\mathrm{DQ},k}(0) = p_{i,0}^{\mathrm{DQ},k-1}(\delta)$ denotes the probability that the downstream queue of link i is empty at the beginning of time interval k, and $p_{i,\ell_i}^{\mathrm{UQ},k}(0) = p_{i,\ell_i}^{\mathrm{UQ},k-1}(\delta)$ denotes the probability that the upstream queue of link i is full at the beginning of time interval k, see Equations (8) and (9).

The physical assumptions of this specification are consistent with vehicle traffic phenomena. The limited free-flow travel time ensures finite vehicle progressions in uncongested conditions. Locating the queue service of the DQ at the downstream end of the link corresponds to the bottleneck nature of (possibly signalized) intersections. Limiting the occupancy of a link by its space capacity, which is captured via the finite capacity queueing framework, allows to capture spillbacks. Furthermore, the proposed model captures the finite dissipation rate of queues through the use of the UQ.

Figure 2 depicts the fundamental diagram that results from this configuration for deterministic arrival and service processes. The slope of the uncongested half equals the free-flow speed $\hat{v}$ that is defined through k^{fwd}, and the slope of the congested half equals the backward wave speed $\hat{w}$ that is defined through k^{bwd}:

- In stationary uncongested conditions with a constant flow q across the link, the number of vehicles in the link is $qk^{fwd}\delta$ and the vehicle density is $\varrho = qk^{fwd}\delta/L$ where L is the link length. This defines the linearly increasing uncongested part of the fundamental diagram with slope $\hat{v} = q/\varrho \propto 1/k^{fwd}$.

- In stationary congested conditions with a constant flow q across the link, the number of backwards traveling spaces in the link is $qk^{bwd}\delta$. Since every space indicates the absence of a vehicle, the vehicle density is $\varrho = \hat{\varrho} - qk^{bwd}\delta/L$, which defines the linearly decreasing congested part of the fundamental diagram with slope $\hat{w} \propto 1/k^{bwd}$.

Algorithm 1 Network simulation

1. set initial queue distributions $\{p^0_{i,n}(0)\}_{n=0}^{\ell_i}$ of DQ and UQ for all links i
2. repeat the following for time intervals $k = 0, 1, \ldots$
   (a) compute boundaries $P(N_i^k > 0)$ and $P(N_i^k < \ell_i)$ of DQ and UQ for all links i according to Equations (16) and (17)
   (b) compute inflows $q_i^{\text{in},k}$ and outflows $q_i^{\text{out},k}$ for all links i according to Equation (4) and (5)
   (c) compute service and arrival rates for all queues according to Equations (10) and (11), accounting for the relations (12)-(15) between the in- and outflows of links and their respective UQs and DQs
   (d) obtain $\{p^{k+1}_{i,n}(0)\}_{n=0}^{\ell_i}$ of DQ and UQ for all links i from Equation (8) and (9)

A variety of deterministic queueing models that account for these effects in one way or another have been proposed in the literature, e.g., Helbing (2003), Bliemer (2007). A simulation-based implementation of the UQ/DQ approach is described in Charypar (2008). The proposed model contributes by providing probabilistic performance measures in an analytical framework.

2.4 Network model

The equations presented in the previous sections are sufficient to define the flow across a linear sequence of links. Algorithm 1 gives an overview of the procedure used to evaluate the network model. It is important to note that although the given phenomenological specifications are constrained to a linear network, the approach carries over straightforwardly to general network topologies, given that an appropriate node model is applied.

If the modeled scenario represents the traffic dynamics across a whole day, it is plausible to start from an empty network, i.e., setting $p^0_{i,0}(0) = 1$ and $p^0_{i,n}(0) = 0$, $n = 1 \ldots \ell_i$ for the DQ and UQ of all links i. If the analysis period does not start at the beginning of a day, initial queue length distributions from a whole-day modeling effort can be used.

Algorithm 1 omits exogeneous demand entries and demand exits for simplicity. These can be accounted for by (i) attaching an infinite capacity link to each demand entry point, (ii) feeding the demand into this link, and (iii) computing physical flow entries into the network by application of the node model downstream of the entry link. Demand exits can be captured by removing a share of the flow transmissions at the exit points, or by adding infinite capacity exit links and allowing for departure turns into those links. Various implementations of this type of network boundary logic can be found in virtually every traffic network model.

In summary, the network model exhibits two important features that render it applicable to real scenarios:

- The model requires as few parameters as the most simple first-order models: link geometry, capacities, maximum velocities, jam densities. This makes it easy to calibrate. Furthermore, its differentiability suggests that efficient optimization-based calibration procedures are applicable.

parameter	value	normalized
vehicle length	5 m	1 slot
link length	100 m	20 slots
max. density $\hat{\varrho}$	200 veh/km	1 veh/slot
time step length	1 s	1 s
free flow velocity $\hat{v}$	36 km/h	2 slot/s
backward wave speed $\hat{w}$	18 km/h	1 slot/s

Table 1: Parameters of test scenario

- The model solution logic resembles the fairly standard process of alternately (i) computing flows from link densities and (ii) computing link densities from flows. This allows for a clear and efficent implementation that exploits the usual decoupling of non-adjacent links within a single time interval.

3 Experiments

In this section, we investigate the performance of the proposed model for a homogeneous link in different congestion regimes and in both dynamic and stationary conditions. The purpose of these experiments is to demonstrate the model's capability of (i) dynamically capturing the build-up, dissipation, and spillback of probabilistic queues on the link, and to (ii) generate a plausible fundamental diagram in stationary conditions. A comparison to the behavior of the KWM in identical conditions is also given. All experiments share the geometrical settings given in Table 1. The second column in this table gives the physical characteristics of the link, and the third column offers a normalized version of these quantities.

3.1 Experiment 1: queue build-up, spillback, and dissipation

This experiment investigates the behavior of the proposed model in dynamic conditions. We assume an initially empty link and an arrival rate that is 0.3 veh/s for the first 500 s and then jumps down to 0.1 veh/s, where it stays for the remaining 500 s. For greater realism, in particular in order to resemble the embedding of the link in a real network, we apply Robertson's recursive platoon dispersion model before feeding the arrivals into the link (Robertson, 1969).

The downstream flow capacity of the link is 0.2 veh/s, which implies that the first half of the demand exceeds the link's bottleneck capacity, whereas the second half can be served by the bottleneck. Drawing from the KWM, one would expect the build-up of a queue, its eventual spillback to the upstream end of the link, and, after 500 s, its (eventually complete) dissipation.

The experimental results are given in Figures 3 and 4. The top row of either Figure shows the results obtained with the probabilistic queueing model, and the bottom row shows the respective results obtained with the KWM. The KWM results are generated using a cell-transmission model (Daganzo, 1994), where the parameters of Table 1 result in the triangular fundamental shown in Figure 2.

Figure 3 displays six diagrams, all of which represent trajectories over time: the first row contains results obtained with the stochastic queueing model, and the second row contains results obtained

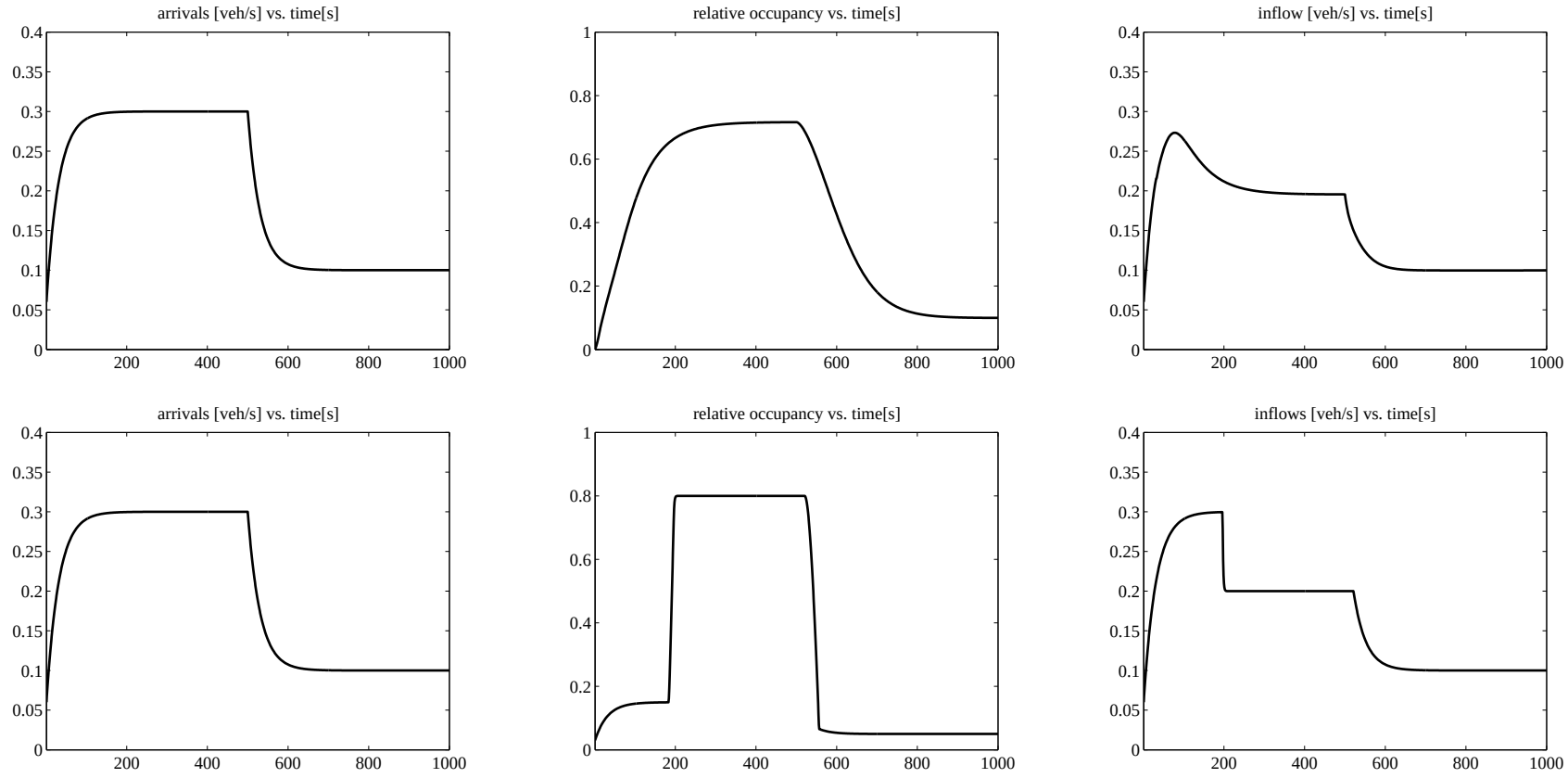

Figure 3: Transient link behavior under changing boundary conditions

with the KWM. The first column shows the upstream flow arrival profile (identical in either case), the second column shows the evolution of the relative occupancy of the link (number of vehicles divided by maximum number of vehicles), and the third column shows the actually realized inflow profile.

The arrivals in the first column of Figure 3 represent the dispersed version of a rectangular demand profile that jumps from 0.3 veh/s to 0.1 veh/s after 500 seconds. We assume the rectangular profile to appear 60 seconds upstream of the considered link and apply a platoon dispersion factor of 0.5, which results in a smoothing factor of approx. 0.032 in Robinson's formula (Robertson, 1969).

The evolution of the relative occupancy and the inflow rate in the second and third column of Figure 3 is in coarse terms similar for the queueing model and the KWM: in either case, more vehicles enter than leave the link during the first 200 s, and hence the occupancy grows. As from second 200, the downstream bottleneck has spilled back to the upstream end of the link, limiting its inflow to the bottleneck capacity (0.2 veh/s) and maintaining a stable and relatively high traffic density on the link. Starting in second 500, the demand drops to half the bottleneck capacity, the queue dissipates, and the occupancy eventually stabilizes again at a relatively low value. Two key differences between the queueing model and the KWM can be identified.

- Both the spillback and its dissipation occur at crisp points in time (i.e., instantaneously) in the KWM, whereas they happen gradually in the stochastic queueing model. This is so because the queueing model captures spillback as a probabilistic event and the respective curves represent *expectations* over distributed occupancies and inflows.

- After the arrival drop in second 500, the expected link occupancy in the stochastic queueing model stabilizes at a higher value (0.1) than in the KWM (0.05). Again, the reason for this is the randomness in the queueing model, which allows for the occurence of downstream queues even in undersaturated conditions, which is not possible in the deterministic KWM.

The three columns in Figure 4 display the temporal evolution of the upstream boundary conditions the link provides, the respective downstream boundary conditions, and its actual outflow. Again, there is qualitative agreement between the stochastic queueing model and the KWM. In the former, the upstream boundary conditions (first column) are given in terms of the *no-spillback probability* that the upstream end of the link is occupied (or blocked) by a vehicle, whereas the KWM captures this effect through a deterministic link supply function. A decrease in the no-spillback probability (i.e., an increase in the spillback probability) is paralleled by a reduced link supply: the no-spillback probability decreases concurrently with the link occupancy and stabilizes after 200 s around 0.65. This is plausible: since the bottleneck capacity is 2/3 of the demand during the first 500 s, 1/3 of the arrivals are rejected. After second 500, the no-spillback probability quickly approaches a value of almost one. This indicates that, although there remains a queue in the link, it does not spill back far enough to affect its inflow.

Finally, the downstream boundary conditions and the resulting outflows (second and third column of Figure 4) are also consistent for both models. The stochastic queueing model captures the downstream boundary in terms of the probability that a vehicle is available at the downstream end of the link, whereas the KWM models this through the deterministic downstream demand of the link. As the link runs full during the first 500 seconds, there is almost always a vehicle available (or ready) to leave the link. Once the demand drops to half of the bottleneck capacity, the availability of a downstream vehicle goes down to 0.5: only every second service offered by the bottleneck is claimed by an available vehicle. The KWM follows the same trend, only that the final drop in demand goes

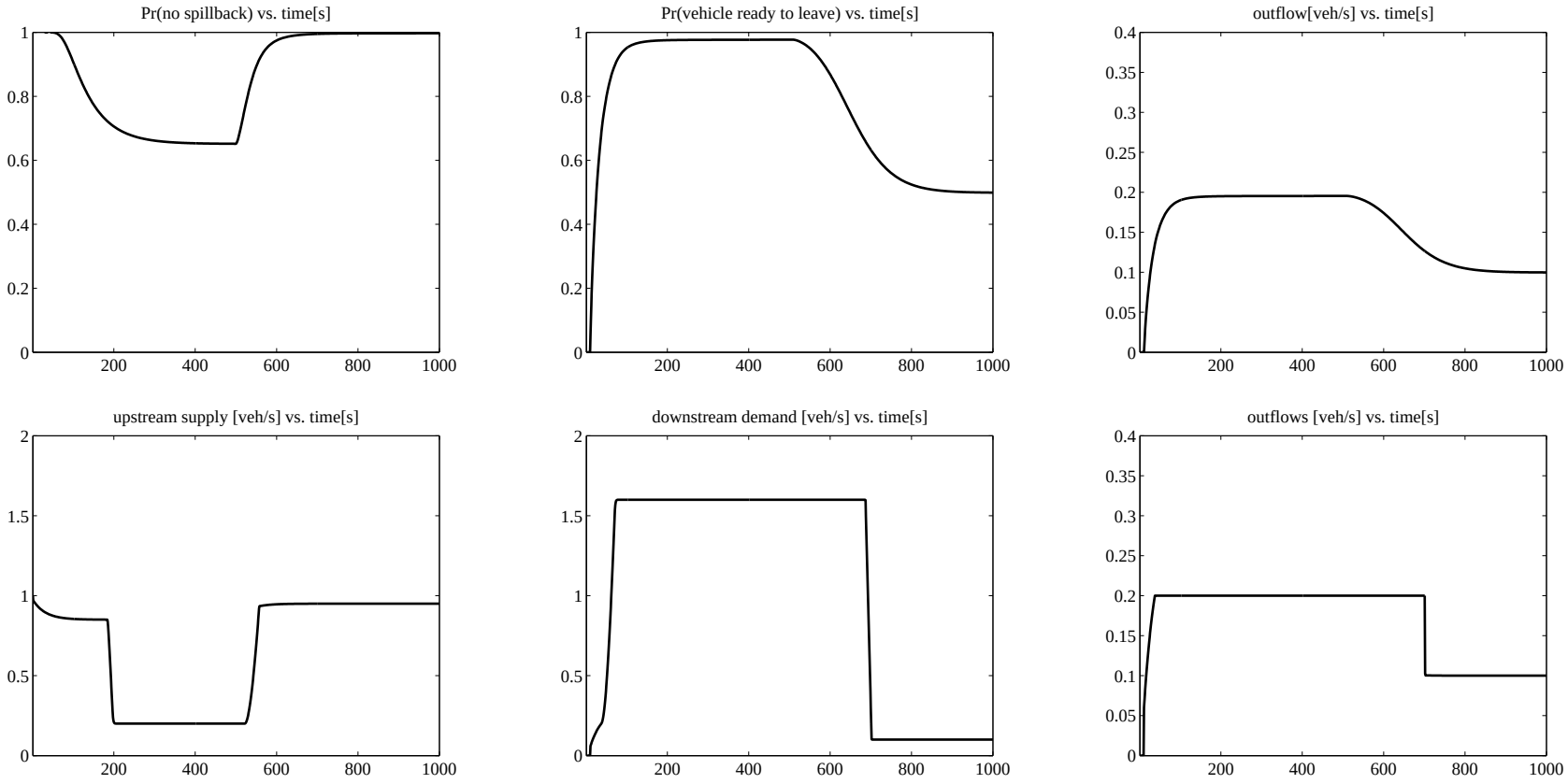

Figure 4: Transient link behavior under changing boundary conditions

further than in the stochastic queueing model. This is, again, a consequence of the occurrence of a stochastic queue in the probabilistic model even in undersaturated conditions, which contains additional, delayed vehicles that are ready to leave the link. Finally, the last column shows that as the link runs full, the outflow rate approaches that of the bottleneck, and as the arrivals drop below the bottleneck capacity in second 500, the outflow follows this trend with some delay, during which the queue in the link dissipates.

It is important to stress that the new model captures all of these effects probabilistically, and hence it allows to assess dynamic traffic conditions with respect to, e.g., their sensitivity to occasional link spillbacks and the resulting network gridlocks. This property is particularly important for short links, where queue spillbacks can quickly reach the upstream intersection. Also noteworthy is that all of these effects are captured by differentiable equations, which makes the model amenable to efficient optimization procedures for, e.g., signal control (Osorio, 2010; Osorio and Bierlaire, 2009b; Osorio and Bierlaire, 2009c) or mathematical formulations of the dynamic traffic assignment problem (Peeta and Ziliaskopoulos, 2001).

3.2 Experiment 2: fundamental diagram

This experiment investigates the behavior of the proposed model in stationary conditions. It does so by creating boundary conditions that in the demand/supply framework of the KWM would reproduce the fundamental diagram. The questions answered by this experiment are (i) if the proposed model has a plausible fundamental diagram and (ii) how this fundamental diagram compares to its deterministic pendant discussed in Section 2.3.3 and shown in Figure 2.

The left (uncongested) and right (congested) half of the fundamental diagram are independently generated. For every point in the uncongested half, the downstream bottleneck capacity is set to a large value, external arrivals are generated at a constant rate, and the system is run until stationarity. The resulting pair of density in the link and flow across the link constitutes one point of the diagram. This experiment is repeated for many different arrival rates between zero and the bottleneck capacity. For every point in the congested half, the downstream bottleneck is set to a particular value, external arrivals are generated at a high rate, and the system is run until stationarity. Again, the resulting pair of density in the link and flow across the link constitutes one point of the diagram. This experiment is repeated for many different bottleneck capacities between zero and the arrival rate.

Figure 5 displays two fundamental diagrams that are generated in this way. Consider first the solid curve. Its slope at low densities approaches the free-flow velocity, and its slope at high densities the backward wave velocity. The curve is concave and reaches its maximum value at a critical density of 0.4 veh/slot, yielding an effective link capacity of 0.5 veh/s. Since the current implementation of the model is fairly prototypical, special situations like zero in- or outflows that constitute limit cases of the System of Equations (8) cannot be treated for numerical reasons; the extreme ends of the fundamental diagram therefore need to be taken with care. In particular, there is a positive flow computed even at maximum density. This is likely to result from numerical imprecisions and needs further investigation.

The only difference between the solid and the dashed fundamental diagram is that the "very large" value chosen for the bottleneck/arrivals when computing the uncongested/congested half of the fundamental diagram is different: in the solid case, it is 0.67 veh/s (this corresponds to the capacity of a triangular fundamental diagram with the same parameters), and in the dashed case

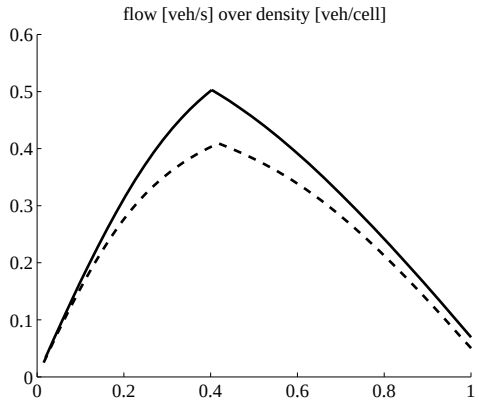

Figure 5: Two fundamental diagrams obtained with the stochastic queueing model

it is 0.5 veh/s. This shows that in the probabilistic model, the link capacity (maximum of the fundamental diagram) is not an invariant quantity – which can be explained by the considerations given in Section 2.1: even in uncongested conditions, the downstream bottleneck capacity takes effect, and even in congested conditions, the upstream arrival rate plays a role. Ultimately, experiments with real data are necessary to assess the physical correctness of the proposed model.

4 Conclusions

This paper presents a dynamic network loading model that resorts to finite capacity queueing theory in order to capture the interactions between upstream and downstream queues (e.g., spillbacks). The method, which builds upon a stationary queueing network model, yields dynamic analytical queue length distributions. The novel dynamic formulation of this model consists of a dynamic link model and a static node model. The stationary probability equations are replaced by a discrete-time closed-form expression for the transient queue length distributions. This expression, which guides the transition of the distributions from one time step to the next, is available in closed form under the reasonable assumption of constant link boundary conditions during a simulation step. No dynamics are introduced into the node model, which maintains the structure of the original stationary model.

Experimental investigations of the proposed model are presented. A comparison with results predicted by the KWM shows that the new model correctly represents the dynamic build-up, spillback, and dissipation of queues. It goes beyond the KWM in that it captures queue lengths and spillbacks probabilistically, which allows for a richer analysis than the deterministic predictions of the KWM. The new model also generates a plausible fundamental diagram, which demonstrates that it captures well the stationary flow/density relationships in both congested and uncongested conditions.

There are various applications of this model. Full dynamic queue length distributions can be used as inputs for route or departure time choice models that capture risk-averse behavior. The analytically tractable form of the stationary model has enabled us in the past to use it to solve traffic control problems using gradient-based optimization algorithms. Since the dynamic formulation preserves the smoothness of the original model, we expect it to be of equal interest for problems that involve derivative-based algorithms, including solution procedures for the dynamic traffic assignment problem.

References

Alfa, A. S. and Neuts, M. F. (1995). Modelling vehicular traffic using the discrete time markovian arrival process, *Transportation Science* **29**(2): 109–117.

Bliemer, M. (2007). Dynamic queuing and spillback in an analytical multiclass dynamic network loading model, *Transportation Research Record* **2029**: 14–21.

Bocharov, P. P., D'Apice, C., Pechinkin, A. V. and Salerno, S. (2004). *Queueing theory*, Modern Probability and Statistics, Brill Academic Publishers, Zeist, The Netherlands, chapter 3, pp. 96–98.

Boel, R. and Mihaylova, L. (2006). A compositional stochastic model for real time freeway traffic simulation, *Transportation Research Part B: Methodological* **40**: 319–334.

Brockfeld, E. and Wagner, P. (2006). Validating microscopic traffic flow models, *Proceedings of the 9th IEEE Intelligent Transportation Systems Conference*, Toronto, Canada, pp. 1604–1608.

Cetin, N., Burri, A. and Nagel, K. (2002). Parallel queue model approach to traffic microsimulations, *Proceedings of the Seventh Swiss Transport Research Conference*, Ascona, Switzerland.

Charypar, D. (2008). Efficient algorithms for the microsimulation of travel behavior in very large scenarios.

Daganzo, C. (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. (1995a). The cell transmission model, part II: network traffic, *Transportation Research Part B* **29**(2): 79–93.

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

Flötteröd, G. and Rohde, J. (2009). Modeling complex intersections with the cell-transmission model, *Technical Report TRANSP-OR 090719*, Ecole Polytechnique Fédérale de Lausanne & Technical University Carolo-Wilhelmina of Braunschweig.

Garber, N. J. and Hoel, L. A. (2002). *Traffic and Highway Engineering*, 3rd edn, Books Cole, Thomson Learning, chapter 6, pp. 204–210.

Greenshields, B. (1935). A study of traffic capacity, *Proceedings of the Annual Meeting of the Highway Research Board*, Vol. 14, pp. 448–477.

Hall, R. W. (2003). *Transportation Queueing*, International series in operations research and management science, Kluwer Academic Publishers, Boston, MA, USA, chapter 5, pp. 113–153.

Heidemann, D. (1994). Queue length and delay distributions at traffic signals, *Transportation Research Part B: Methodological* **28**(5): 377–389.

Heidemann, D. (1996). A queueing theory approach to speed-flow-density relationships, *Proceedings of the* 13^{th} *International Symposium on Transportation and Traffic Theory*, Lyon, France, pp. 103–118.

Heidemann, D. (2001). A queueing theory model of nonstationary traffic flow, *Transportation Science* **35**(4): 405–412.

Heidemann, D. and Wegmann, H. (1997). Queueing at unsignalized intersections, *Transportation Research Part B: Methodological* **31**(3): 239–263.

Helbing, D. (2003). A section-based queuing-theoretical model for congestion and travel time analysis in networks, *Journal of Physics A: Mathematical and General* **36**: L593–L598.

Hilliges, M. and Weidlich, W. (1995). A phenomenological model for dynamic traffic flow in networks, *Transportation Research Part B* **29**(6): 407–431.

Hoogendoorn, S. and Bovy, P. (2001). State-of-the-art of vehicular traffic flow modelling, *Proceedings of the Institution of Mechanical Engineers. Part I: Journal of Systems and Control Engineering* **215**(4): 283–303.

Jain, R. and Smith, J. M. (1997). Modeling vehicular traffic flow using M/G/C/C state dependent queueing models, *Transportation science* **31**(4): 324–336.

Kerbache, L. and Smith, J. M. (2000). Multi-objective routing within large scale facilities using open finite queueing networks, *European Journal of Operational Research* **121**(1): 105–123.

Kotsialos, A., Papageorgiou, M., Diakaki, C., Pavlis, Y. and Middelham, F. (2002). Traffic flow modeling of large-scale motorway networks using the macroscopic modeling tool METANET, *IEEE Transactions on Intelligent Transportation Systems* **3**(4): 282–292.

Lebacque, J. (1996). The Godunov scheme and what it means for first order traffic flow models, *in* J.-B. Lesort (ed.), *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, Pergamon, Lyon, France.

Lebacque, J. (2005). First–order macroscopic traffic flow models: intersection modeling, network modeling, *in* H. Mahmassani (ed.), *Proceedings of the 16th International Symposium on Transportation and Traffic Theory*, Elsevier, Maryland, USA, pp. 365–386.

Lighthill, M. and Witham, J. (1955). On kinematic waves II. a theory of traffic flow on long crowded roads, *Proceedings of the Royal Society A* **229**: 317–345.

Morse, P. (1958). *Queues, inventories and maintenance. The analysis of operational systems with variable demand and supply*, Wiley, New York.

Newell, G. F. (1979). *Approximate behavior of tandem queues*, Vol. 171 of *Lecture notes in economics and mathematical systems*, Springer-Verlag, Berlin.

Oliver, R. M. and Bisbee, E. F. (1962). Queuing for gaps in high flow traffic streams, *Operations Research* **10**(1): 105–114.

Osorio, C. (2010). *Mitigating network congestion: analytical models, optimization methods and their applications*, PhD thesis, Ecole Polytechnique Fédérale de Lausanne.

Osorio, C. and Bierlaire, M. (2009a). An analytic finite capacity queueing network model capturing the propagation of congestion and blocking, *European Journal Of Operational Research* **196**(3): 996–1007.

Osorio, C. and Bierlaire, M. (2009b). A multi-model algorithm for the optimization of congested networks, *Proceedings of the European Transport Conference (ETC)*, Noordwijkerhout, The Netherlands.

Osorio, C. and Bierlaire, M. (2009c). A surrogate model for traffic optimization of congested networks: an analytic queueing network approach, *Technical Report 090825*, Transport and Mobility Laboratory, ENAC, Ecole Polytechnique Fédérale de Lausanne.

Pan, T., Sumalee, A., Zhong, R. and Uno, N. (2010). The stochastic cell transission model considering spatial and temporal correlations for traffic states prediction, *Proceedings of the 3rd International Symposium on Dynamic Traffic Assignment*, Takayama, Japan.

Pandawi, S. and Dia, H. (2005). Comparative evaluation of microscopic car-following behavior, *IEEE Transactions on Intelligent Transportation System* **6**(3): 314–325.

Payne, H. (1971). Models of freeway traffic and control, *Mathematical Models of Public Systems*, Vol. 1, Simulation Council, La Jolla, CA, USA, pp. 51–61.

Peeta, S. and Ziliaskopoulos, A. (2001). Foundations of dynamic traffic assignment: the past, the present and the future, *Networks and Spatial Economics* **1**(3/4): 233–265.

Peterson, M. D., Bertsimas, D. J. and Odoni, A. R. (1995). Models and algorithms for transient queueing congestion at airports, *Management Science* **41**(8): 1279–1295.

Richards, P. (1956). Shock waves on highways, *Operations Research* **4**: 42–51.

Robertson, D. (1969). TRANSYT: a traffic network study tool, *Technical Report Rep. LR 253*, Road Res. Lab., London, England.

Sumalee, A., Zhong, R., Pan, T., Iryo, T. and Lam, W. (2010). Stochastic cell transission model for traffic network with demand and supply uncertainties, *Proceedings of the 3rd International Symposium on Dynamic Traffic Assignment*, Takayama, Japan.

Sumalee, A., Zhong, R., Szeto, W. and Pan, T. (2009). Stochastic cell transmission model: traffic state modeling under uncertainties, *Proceedings of the 88. Annual Meeting of the Transportation Research Board*, Washington, DC, USA.

Tampere, C., Corthout, R., Cattrysse, D. and Immers, L. (forthcoming). A generic class of first order node models for dynamic macroscopic simulation of traffic flows, *Transportation Research Part B: Methodological* **xx**: xx–xx.

Tanner, J. C. (1962). A theoretical analysis of delays at an uncontrolled intersection, *Biometrika* **49**: 163–170.

Van Woensel, T. and Vandaele, N. (2007). Modelling traffic flows with queueing models: A review, *Asia-Pacific Journal of Operational Research* **24**(4): 1–27.

Viti, F. (2006). *The Dynamics and the Uncertainty of Delays at Signals*, PhD thesis, Delft University of Technology. TRAIL Thesis Series, T2006/7.

Yeo, G. F. and Weesakul, B. (1964). Delays to road traffic at an intersection, *Journal of Applied Probability* **1**(2): 297–310.