A differentiable dynamic network loading model that yields queue length distributions and accounts for spillback

Carolina Osorio * Gunnar Flötteröd * Michel Bierlaire *

Abstract

We derive a dynamic network loading model that yields queue length distributions, properly accounts for spillback, 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 elementary situations under different congestion regimes, comprising queue build-up, spillback, and dissipation.

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. Here, only passenger vehicle traffic is considered, the DNL problem is to capture the traffic flow dynamics on the road network.

In this article, we concentrate on macroscopic models, and we do so for the usual reasons: low number of model parameters to be calibrated, good computational performance, mathematical tractability. For reviews on microsimulation-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 density and velocity are uniquely related and flow is a function of either density or velocity (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.

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

Models of the 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 Lebacque (1996) and further developed in Lebacque (2005) provides a comprehensive theoretical basis for basic first-order node models, where flow interactions result from limited inflow capacities of the downstream links. Recently, this framework has been supplemented with richer features such as the modeling of conflicts within the node (Flötteröd and Rohde, 2009; Tampere et al., 2010).

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. To the best of our knowledge, no consistent generalization of this theory for probabilistic traffic conditions has been put forward in an analytical framework.

The typical course of action is to resort to microsimulations, which account for arbitrary sources of stochasticity but only generate realizations of the underlying distributions. A notable exception is the work of Sumalee et al. (2008) on a stochastic version of the cell-transmission model (Daganzo, 1994; Daganzo, 1995a). Our approach differs from this work in that it (i) exploits closed-from transient results from queueing theory, (ii) provides the additional benefit of a closed-form expression of the system's stationary state, and (iii) consists of one integrated set of smooth equations, whereas Sumalee et al. (2008) deploy a switching logic between multiple linear models.

The model proposed in this work is based on finite capacity queueing theory, which provides a variety of analytical probabilistic expressions for the evolution of (e.g., vehicular) queues. A review of urban traffic models based on queueing theory is given in Osorio (2010). Given the complexity of deriving tractable analytical expressions for transient distributions, existing methods focus on stationary distributions. Most methods evaluate the stationary distributions of a given lane at a given intersection. Stationary distributions for several lanes and several links have also been proposed (Osorio and Bierlaire, 2009b).

In this article, we extend the analytical stationary model of Osorio and Bierlaire (2009b) to an analytical transient model, deriving dynamic queue length distributions. The original 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 in detail the temporal build-up and dissipation of queues. Adding dynamics to this type of model is a novel undertaking, and we conceive this work to be the first consistent representation of queue length distributions in the dynamic network loading problem.

2 Model

Most of this text treats the probabilistic modeling of traffic flow on a homogeneous road segment (link). The development of a node model is the topic of ongoing research; the model of Osorio and Bierlaire (2009a) constitutes the starting point of this work.

2.1 Relation between the KWM and finite capacity queues

As usual, we represent a homogeneous link segment as a set of queues, with the main innovation being that the queueing model describes a distribution of the queue length through analytical equations.

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 leave 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 (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

- in congested conditions, increasing the upstream demand does not affect the flow, and
- in uncongested conditions, increasing the downstream supply does not affect the flow.

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 essentially 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 in all traffic states.

Interestingly, these considerations suggest that every probabilistic version of the KWM will also violate the invariance principle. 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

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

where $p(\cdot)$ represents the respective probability density functions. The possibility o f inconsistencies with the KWM results from the fact that the above expectation can mix free-flow and congested traffic conditions.

This indicates that the proposed model is not a stochastic version of the KWM. Further investigating the links between these model classes is an interesting topic of future research, currently this link only serves as a conceptual guideline when developing the new model.

2.2 Finite capacity queueing model

We build upon the urban traffic model formulated in 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 segments (i.e., spillbacks), and to consider congested scenarios where traffic demand may exceed supply. In queueing theory terms, this corresponds to a traffic intensity that may exceed one. This is the main distinction between classical queueing theory and finite capacity queueing theory.

The key feature of the model is the description of the interactions among the different queues. 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 is given by the *blocking probability* of a queue.

All distributional assumptions and approximations of the model of Osorio and Bierlaire (2009a) are preserved in this framework. A detailed discussion appears 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.

The stationary model consists of a system of nonlinear equations. It is formulated based on a set of exogenous parameters that capture the network topology, the total demand, and the turning probabilities. A set of endogenous variables describes the traffic interactions, including the spillback probabilities and the rates at which a spillback diffuses.

2.3 Dynamic model

2.3.1 Queueing model

The stationary model we start from 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:

i	queue index;
k	time interval index;
δ	time step length;
$p_{i,n}^k(t)$	transient probability that queue i is of length n at time t, where t is in time interval k;
$s_{i,n}^k$	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 provided to queue i during time interval k;
λ_i^k	arrival rate of queue i during time interval k;
ℓ_i	upper bound of the queue length of queue i (referred to in queueing theory as the *queue capacity*).

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*:

$$\rho_i^k = \frac{\lambda_i^k}{\hat{\mu}_i^k}. \tag{2}$$

As described in Section 2.2, 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:

$$s_{i,n}^k = \frac{1-\rho_i^k}{1-(\rho_i^k)^{\ell_i+1}}(\rho_i^k)^n. \tag{3}$$

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

The transient probabilities for a given queue i and a given time interval k are given by: $\forall n = 0, 1, ..., \ell_i, \forall t \in [0, \delta]$

$$\begin{cases} p_{i,n}^k(t) = s_{i,n}^k + (\rho_i^k)^{\frac{n}{2}} \sum_{j=1}^{\ell_i} C_{i,j}^k \left\{ \sin \frac{jn\pi}{\ell_i+1} - \sqrt{\rho_i^k} \sin \frac{j(n+1)\pi}{\ell_i+1} \right\} e^{\tau_{i,j}^k t} & \text{(4a)} \\ \tau_{i,s}^k = \lambda_i^k + \hat{\mu}_i^k - 2\sqrt{\lambda_i^k \hat{\mu}_i^k} \cos \frac{s\pi}{\ell_i+1}, & \text{(4b)} \end{cases}$$

where the coefficients $\{C_{i,j}^k\}_j$ are chosen to fit the initial values of the transient distribution p at the beginning of the time interval, as follows:

$$p_{i,n}^k(0) = p_{i,n}^{k-1}(\delta). \tag{5}$$

These boundary conditions (Equation (5)) maintain the temporal continuity of the queue length distributions. For details on the derivation of the transient probabilities, see Morse (1958).

2.3.2 Node model

We model a set of links in series (i.e., a set of queues in a tandem topology). Vehicles arrive to the first link, travel along all links and leave the network at the last link, i.e., vehicles enter the network through the first queue only and leave the network from the last queue only. This formulation can be extended to allow for arbitrary link topologies as well as a 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.

To formulate the node model we introduce the following notation:

$q_i^{\text{in},k}$	inflow to queue i during time interval k (in vehicles per time unit);
$q_i^{\text{out},k}$	outflow from queue i during time interval k (in vehicles per time unit);
μ_i^k	exogenous supply provided to queue i during time interval k (in vehicles per time unit);
N_i	total number of vehicles in queue i.

Flow conservation defines the inflow of a given queue as the outflow of its upstream queue:

$$\forall i > 1, \; q_i^{\text{in},k} = q_{i-1}^{\text{out},k}, \tag{6}$$

The outflow of a queue is given by queueing theory as:

$$q_i^{\text{out},k} = \hat{\mu}_i^k P^{k-1}(N_i > 0), \tag{7}$$

where $\hat{\mu}_i^k$ is the service rate (downstream supply) of the queue and $P^{k-1}(N_i > 0)$ denotes the probability that the queue is not empty, i.e., that there is at least one vehicle queueing at the downstream end of the link.

The service rate of a queue is given by:

$$\hat{\mu}_i = \mu_i^k(1 - P^{k-1}(N_{i+1} = \ell_{i+1})). \tag{8}$$

That is, the service rate of a queue is determined by an exogenous supply provided to that queue (which may depend on link properties (e.g. length, free-flow speed) and intersection attributes (e.g. signal plans, ranking of traffic streams)). This equation states that supply is provided as long as the downstream queue does not spill back.

The inflow to a queue and its arrival rate are linked as follows:

$$\lambda_i^k = q_i^{\text{in},k}/(1 - P^{k-1}(N_i = \ell_i)) \tag{9}$$

where $P^{k-1}(N_i = \ell_i)$ represents the probability that queue i is full at the end of time interval $k-1$ (i.e., spillback or blocking probability). This probability is also known in finite capacity queueing theory as the *loss probability*. For such queues, the arrivals that arise while the queue is full are considered to be "lost". In traffic flow modeling, the "loss" of a vehicle corresponds to its disability to enter the downstream link, and hence "lost" vehicles must be stored upstream to maintain mass conservation. Thus the arrival rate, λ_i^k, represents the potential demand that arises at queue i, and the inflow, $q_i^{\text{in},k}$, captures the demand that actually enters the link.

This spillback probability and the probability that queue i is not empty, $P^{k-1}(N_i > 0)$, are obtained by evaluating the transient probability distributions as follows.

$$P^k(N_i = \ell_i) = p_{i,\ell_i}^k(\delta), \tag{10}$$

$$P^k(N_i > 0) = 1 - p_{i,0}^k(\delta), \tag{11}$$

where δ is the time step length.

2.3.3 Link model

The node model (Equations (6)-(9)) couples the dynamics of upstream demand and downstream supply through two parameters. The first is the probability that there are vehicles at the end of a link ready to proceed to a downstream link, $P^k(N_i > 0)$. The second is the probability $P^k(N_i = \ell_i)$ that a link spills back.

In order to approximate 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).

The role of the downstream queue is to capture the downstream dynamics of the link in order to approximate $P^k(N_i > 0)$. This downstream queue considers all vehicles that are at the downstream end of the link waiting to proceed to a downstream link, i.e., all vehicles that are in the physical queue of the link. This queue is depicted in Figure 1.

To accurately represent the length of this queue we account for the time needed for a vehicle that enters the link to reach the physical queue. As is depicted in Figure 1, the inflow to DQ, $q^{in}(k - k^{fwd})$, is equal to the inflow of the link lagged by a fixed factor, k^{fwd}, which represents the time needed for the vehicle to reach the queue. This imposes a lower bound on the free-flow travel time, which ensures finite vehicle progressions in uncongested conditions. Furthermore, the outflow of DQ corresponds to the link outflow, $q^{in}(k)$.

The role of the upstream queue is to capture the upstream dynamics of the link in order to derive the spillback probability $P^k(N_i = \ell_i)$. This queue therefore captures the finite dissipation rate of queues. In other words, upon the departure of a vehicle from a link it accounts for the time needed for this newly available space to reach the end of the link.

This is achieved by setting the outflow of UQ, $q^{out}(k - k^{bwd})$, equal to that of DQ 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. Additionally, the inflow of the UQ is the link inflow, $q^{in}(k)$. 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.

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

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

A variety of deterministic queueing models that account for the same effects have been proposed in the literature, e.g., Helbing (2003), Bliemer (2007). The proposed model innovates by providing a probabilistic performance measures.

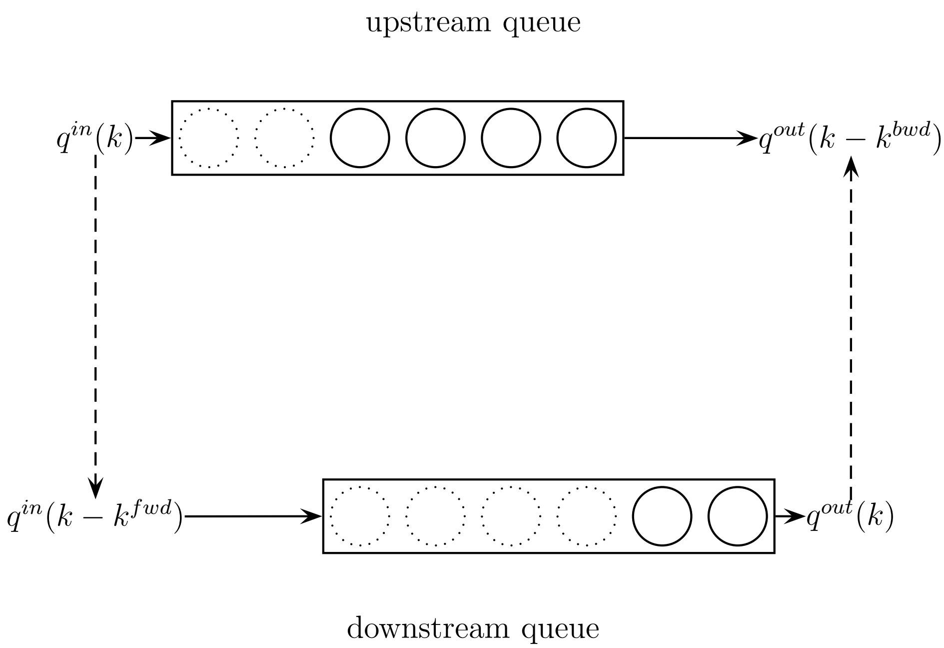

Figure 1: Link modeled with two time-shifted queues

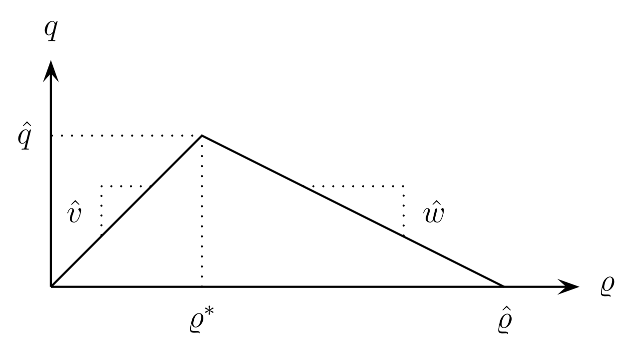

Figure 2: Fundamental diagram for deterministic double-queue model

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

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 buildup, dissipation, and spillback of probabilistic queues on the link, and to (ii) generate a plausible fundamental diagram in stationary conditions. 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 buildup, 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. 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 buildup of a queue, its eventual spillback to the upstream end of the link, and, after 500 s, its (eventually complete) dissipation.

Figure 3 displays six diagrams, all of which represent trajectories over time: the first row contains, from left to right, the average density on the link (in vehicles per slot), the probability that the upstream end of the link is occupied by a vehicle, and the probability that a vehicle is present at the downstream end of the link. The second row contains, also from left to right, the external arrival rate, the link inflow rate, and the link outflow rate, all in vehicles per second.

The evolution of the density (first figure of first row) is intuitively plausible: during the first 200 s, more vehicles enter than leave the link, and hence the density grows. As from second 200, the downstream bottleneck has spilled back to the upstream end of the link, limiting its inflow and maintaining a stable and relatively high traffic density on the link. As from second 500, the demand drops to half the bottleneck capacity, and the queue dissipates. The occupancy then stabilizes again around 0.1 veh/slot, which is substantially above what one would expect in a deterministic supply regime of the KWM (where for 0.1 veh/s inflow and a free flow travel time of 10 s only 1 vehicle, i.e., 0.05 veh/slot should remain in the link). This is due to the persistence of a stochastic queue in the link even in undersaturated conditions.

The second and third figure in the first row represent the probability that the upstream end of the link is occupied (or blocked) by a vehicle (i.e., the spillback probability) and the probability that

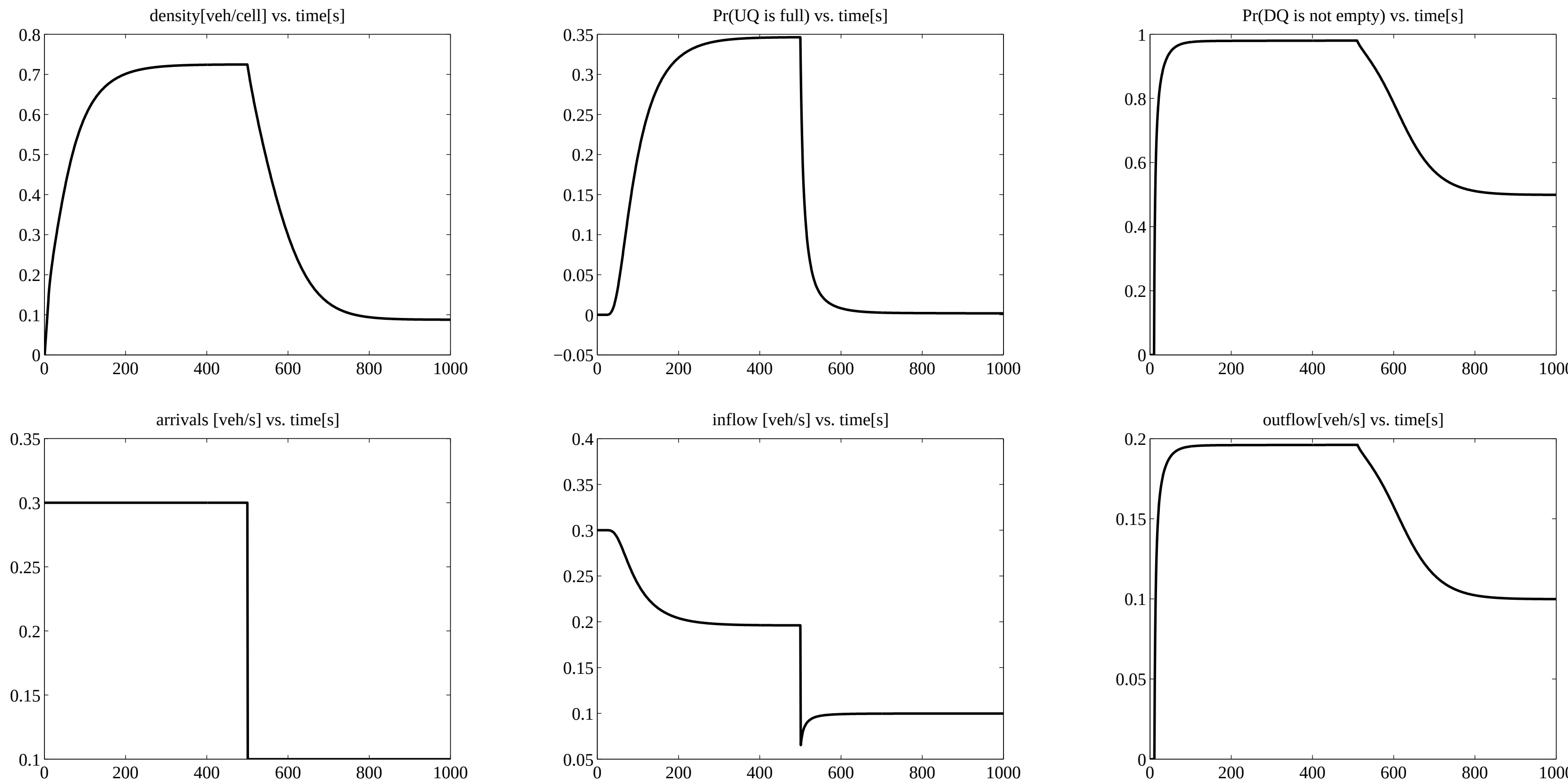

Figure 3: Transient link behavior under changing boundary conditions

a vehicle is available at the downstream end of the link, respectively. The spillback probability rises concurrently with the link occupancy and stabilizes after 200 s around 0.35. 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 blocking probability quickly approaches a value of almost zero. This indicates that, although there remains a queue in the link, it does not spill back far enough to affect its inflow.

The probability that the downstream queue is not empty (figure three) is also intuitively consistent: 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 first and second figure in the lower row show the external arrivals and the link inflow, respectively. As the link runs full during the first 500 s, more and more vehicles are rejected, and eventually only as much inflow as allowed by the downstream bottleneck (0.2 veh/s) is enabled. When the demand jumps from 0.3 veh/s down to 0.1 veh/s, the inflow undershoots somewhat before stabilizing at the arrival rate. This is so because the link inflow is the product of arrival rate and blocking probability. While the arrivals are reduced immediately, the blocking probability decreases only gradually, cf. figure two in row one. Finally, the last figure in the third row shows that as the link runs full, the outflow rate approaches that of the bottleneck, and as the demand goes down after second 300, the outflow also goes down to the arrival rate.

It is important to stress that the 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 spill-backs 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 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 on the link and flow across the link constitutes one point of the diagram. This experiment is repeated for many different arrival rates between almost 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 on the link and flow across the link constitutes one point of the diagram. This experiment is repeated for many different bottleneck capacities between almost zero and the

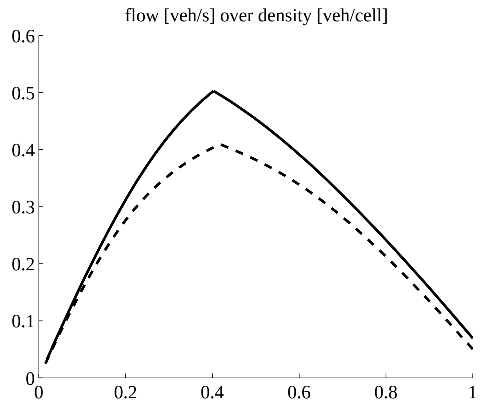

Figure 4: Fundamental diagrams in stationary conditions

arrival rate.

Figure 4 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 (4) 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 arrivals (bottleneck) when computing the uncongested (congested) half of the fundamental diagram is different: in the solid case, it is 0.67 veh/s (corresponds to the capacity of a triangular fundamental diagram with the same parameters), and in the dashed case 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. Clearly, experiments with real data are necessary to assess the physical correctness of the proposed model.

4 Conclusions

We present a dynamic network loading model that yields queue length distributions, properly accounts for spillback, and maintains a differentiable mapping from the dynamic demand on the dynamic queue lengths. The approach builds upon an existing stationary queueing network model based on finite capacity queueing theory. The original model is specified in terms of a set of differentiable equations, which in the proposed model are carried over to a set of equally smooth difference equations.

Essentially, the original stationary model we start from 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-link models. The novel dynamic formulation of this model consists of a dynamic link model and a static node model. The global balance 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. These show that the model correctly represents the dynamic build-up, spillback, and dissipation of queues. Furthermore, it generates a plausible fundamental diagram.

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

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.

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.

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.

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

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.

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.

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.

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.

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

Sumalee, A., Zhong, R., Szeto, W. and Pan, T. (2008). Stochastic cell transmission model under demand and supply uncertainties, *Proceedings of the International Symposium on Dynamic Traffic Assignment*, Leuven, Belgium.

Tampere, C., Courthout, R., Viti, F. and Cattrysse, D. (2010). Link transmission model: an efficient dynamic network loading algorithm with realistic node and link behaviour, *Seventh Triennal Symposium on Transportation Analysis*, Trømso, Norway.