Stochastic network link transmission model

G. Flötteröd * C. Osorio †

April 20, 2017

Abstract

This article considers the stochastic modeling of vehicular network flows, including the analytical approximation of joint queue-length distributions. The article presents two main methodological contributions. First, it proposes a tractable network model for finite space capacity Markovian queueing networks. This methodology decomposes a general topology queueing network into a set of overlapping subnetworks and approximates the transient joint queue-length distribution of each subnetwork. The subnetwork overlap allows to approximate stochastic dependencies across multiple subnetworks with a complexity that is linear in the number of subnetworks. Additionally, the network model maintains mutually consistent overlapping subnetwork distributions. Second, a stochastic network link transmission model (SLTM) is formulated that builds on the proposed queueing network decomposition and on the stochastic single-link model of Osorio and Flötteröd (2015). The SLTM represents each direction of a road and each road intersection as one queueing subnetwork. Three experiments are presented. First, the analytical approximations of the queueing-theoretical model are validated against simulation-based estimates. An experiment with intricate traffic dynamics and multi-modal joint distributions is studied. The analytical model captures most dependency structure and approximates well the simulated network dynamics and joint distributions. Even for the considered simple network, which consists of only eight links, the proposed subnetwork decomposition yields significant gains in computational efficiency: It uses less than 0.0025 % of the memory that is required by the use of a full network model. Second and third, the proposed SLTM is illustrated with a linear test network adopted from the literature and a more general topology network containing a diverge node and a merge node. Time-dependent probabilistic performance measures (occupancy uncertainty bands, spillback probabilities) are presented and discussed.

1 Introduction

This article develops a stochastic network link transmission model (SLTM). Given stochastic network inflows, outflows, and between-link flow transitions, the model (i) describes the state distribution of each link, comprising the joint distribution of the up- and downstream boundary conditions modulating its in- and outflows and (ii) approximates the joint state distribution of multiple links that exchange stochastic dynamic flows.

The network model builds upon an existing model of the transient (i.e. time-dependent) joint distribution of a single homogeneous link's up- and downstream boundary conditions (Osorio and Flötteröd; 2015). This link model is a queueing-theoretical stochastic reformulation of the link transmission model (LTM) of Yperman et al. (2006), which constitutes an operational formulation of Newell's simplified theory of (deterministic) kinematic waves (Newell; 1993). The LTM has received recent attention as a computationally efficient network loading model (Himpe et al.; 2016; Raadsen et al.; 2016). The link model of Osorio and Flötteröd (2015) captures the stochastic flow dynamics within a link through a system of four finite space capacity queues with lagged flows.

The present article contributes to the fields of analytical transient finite capacity queueing network modeling and of vehicular traffic flow modeling.

First, it formulates an analytical tractable approximation of transient joint queue-length distributions in Markovian finite space capacity queueing networks. It does so by decomposing a queueing network topology into a

*KTH Royal Institute of Technology, Department of Transport Science, 11428 Stockholm, Sweden, gunnar.floetteroed@abe.kth.se (corresponding author)

†Massachusetts Institute of Technology (MIT), Department of Civil & Environmental Engineering, Cambridge, MA 02139, USA, osorioc@mit.edu

set of overlapping, non-disjoint subnetworks. This decomposition allows to address the curse of dimensionality. A tractable analytical approximation of the transient joint queue-length distribution of each subnetwork is proposed. It is proven that for queues that belong to multiple subnetworks, the model maintains consistent marginal distributions of the common queues.

Second, this queueing approximation is used to formulate the SLTM. The starting point of this effort is the stochastic single-link model of Osorio and Flötteröd (2015). A network of such links induces a network topology of all queues contained in these links. The queueing theoretical subnetwork decomposition is applied to the resulting queueing network, where (i) all queues within a link constitute one subnetwork and (ii) all queues being adjacent to a node (intersection in the road network) constitute one subnetwork. The queue overlap of link and node subnetworks is the key ingredient enabling the approximation of network-wide stochastic dependencies.

The remainder of this introduction summarizes the state-of-the-art in the two relevant fields of queueing theory and vehicular traffic flow modeling. Section 2 then presents the new queueing network model. It is formulated in Section 2.1 and experimentally validated in Section 2.2. This queueing network model is then used in Section 3 to formulate the proposed network SLTM. The model is formulated in Section 3.1, its numerical solution is discussed in Section 3.2, and its concrete specification and dynamics are illustrated in Section 3.3. Section 4 summarizes the main findings of this work and identifies several important future research topics.

Queueing network analysis

Consistently with queueing-theoretical terminology, the notion of "capacity" refers to "space capacity" throughout this section. The use of finite capacity queues allows to set an upper bound on the queue-size. This accounts for finite physical space capacity and the possible occurrence of spillback into upstream queues. Finite capacity queueing theory does, however, not concern itself with the geometry of the queueing system, it merely considers the number of spaces available in the system and the number of "jobs" (here, vehicles) currently located therein.

The analytical modeling of queueing networks has mostly focused on the *stationary* analysis of systems with *infinite capacity* queues, and more specifically on product-form networks as described in the seminal papers of Jackson (1963, 1957); Baskett et al. (1975). Infinite capacity is a strong assumption for a variety of space-constrained congested networks because it neglects important between-queue dependencies, which are in particular due to blocking phenomena and suggest a non-product form joint distribution. Works such as Odoni and Roth (1983) highlight the importance of carrying out a transient analysis and the inadequacy of using stationary metrics to approximate transients.

For Markovian finite capacity queueing networks (FCQNs), the stationary joint queue-length distribution can be obtained by solving the global balance equations (Stewart; 2000). Other exact numerical methods have been proposed for simple Markovian FCQNs, e.g. with two or three queues in tandem topologies (Grassmann and Derkic; 2000; Akyildiz and von Brand; 1994; Balsamo and Donatiello; 1989; Langaris and Conolly; 1984; Latouche and Neuts; 1980; Konheim and Reiser; 1978, 1976).

For non-product form networks, a major challenge in approximating the joint distribution is its dimensionality. Thus, the most common analytical approach remains that of approximating stationary marginal distributions. Osorio and Bierlaire (2009) provide a review of decomposition techniques that reduce the dimensionality (and hence the computational complexity) by approximating lower-dimensional marginals. The scalable family of aggregation-disaggregation techniques describes the state of the network aggregately (in terms of a reduced state space), while ensuring consistency with disaggregate marginals (e.g. Schweitzer; 1991). A tractable instance of this family for urban transportation networks is given by Osorio and Wang (2017).

Transient techniques have received less attention; this is arguably due to the analytical complexity involved in their analysis. Reviews of transient analysis of queueing models are provided by Kaczynski et al. (2012); Griffiths et al. (2008). For Markovian FCQNs, the transient joint queue-length distribution can be obtained by solving a system of linear first-order ordinary differential equations (ODEs). Closed-form expressions are, to the best of our knowledge, limited to a single M/M/1/K queue (Morse; 1958; Sharma and Gupta; 1982) or a single M/M/2/K queue (Sharma and Shobha; 1988). Numerous exact numerical techniques have been developed (for reviews, see Stewart; 1994, 2009). Although the formulation of the problem as a system of ODEs allows for a variety of numerical ODE techniques to be used, dimensionality remains a major challenge.

Transient decomposition techniques have typically assumed infinite capacity queues (e.g. McCalla and Whitt; 2002; Whitt; 1999; Peterson et al.; 1995a; Odoni and Roth; 1983). Transient decomposition methods for FCQNs have received little attention due to the complexity of providing a tractable analytical description of the temporal between-queue dependencies. A transient and tractable aggregation-disaggregation technique is given by Osorio and Yamani (Forthcoming). Overall, there is currently a lack of analytical transient techniques for

Markovian FCQNs that account for spatial-temporal dependencies, and even more a lack of tractable techniques. This article presents a tractable analytical approximation model of transient multivariate queue-length distributions within a Markovian FCQN.

Vehicular traffic network analysis

The proposed general-purpose queueing-theoretic model is used to formulate a stochastic network model for road traffic that is rooted in mainstream deterministic traffic flow theory. In the broader field of transportation (all modes considered), few queueing-theoretical analytical probabilistic and transient techniques have been developed; see Heidemann (2001); Peterson et al. (1995b) for a single queue and Osorio et al. (2011); Osorio and Flötteröd (2015); Gupta (2011); Peterson et al. (1995a); Odoni and Roth (1983) for networks of queues.

The kinematic wave model (KWM; Lighthill and Witham; 1955; Richards; 1956) is still the mainstay of analytical traffic flow modeling; the previously discussed LTM is consistent with the KWM. Osorio and Flötteröd (2015) propose, in further development of Osorio et al. (2011), a queueing-theoretical stochastic reformulation of the LTM for a single link. Their model captures stochastic link in- and outflows and the resulting stochastic vehicle distribution. Other stochastic link models rely on stochastic cell-transmision models that require a cell-discretization of the link (Boel and Mihaylova; 2006; Sumalee et al.; 2011; Jabari and Liu; 2012).

The so far existing literature on stochastic Newell-type models considers homogeneous road segments but no network topologies. This is the case for the queueing-theoretical model of Osorio and Flötteröd (2015), for the class of stochastic solutions to the KWM with a stochastic initial density profile discussed by Laval and Chilukuri (2014), as well as for the stochastic instances of Newell's three-detector problem formulated by Laval et al. (2012) and Deng et al. (2013). The present article contributes by embedding the stochastic link model of Osorio and Flötteröd (2015) in a network topology.

The existing KWM-consistent node (i.e. intersection) models, which are necessary to model network flows, are deterministic, meaning that they represent (dynamic) space-time average conditions but no additional stochastic information (e.g., Daganzo; 1995b; Lebacque; 1996; Lebacque and Khoshyaran; 2005; Tampere et al.; 2011; Flötteröd and Rohde; 2011; Corthout et al.; 2012; Smits et al.; 2015). The present article develops an SLTM for networks that accommodates many possible stochastic instances of such node models and illustrates this capability through the specification of concrete linear, diverge, and merge node models within the SLTM framework.

In the kinetic approach to stochastic traffic flow modeling, a probabilistic description of individual-vehicle interactions is adopted. This model is then solved in the form of dynamic equations for mean values and variances of aggregate traffic characteristics (e.g. Tampere et al.; 2003). Operational constraints often lead to the simplifying assumption that the states of interacting vehicles are stochastically independent. Nelson and Kumar (2006) discuss the implications of omitting such dependencies. Kinetic models appear as of now too complex to account for realistic dependency structures in non-trivial networks (Helbing; 2001). Such dependencies are captured in the model of the present article.

2 Queueing network model

This section presents the queueing theoretical foundation of the proposed road network SLTM. Section 2.1 formulates the queueing network model, and Section 2.2 presents a simulation-based validation. The material of this section constitutes a stand-alone *queueing* network model. However, all concrete modeling choices and approximations made serve the purpose of facilitating the development of a *road* network SLTM in the subsequent Section 3.

2.1 Model formulation

2.1.1 Full network dynamics and subnetwork decomposition

Consider a network of queues in an arbitrary topology. The queueing network is represented by an undirected and connected graph $G(\mathcal{V}, \mathcal{E})$, where the vertex set $\mathcal{V}$ represents the queues and the edge set $\mathcal{E}$ is such that two queues are connected with an undirected edge if there exists an event that depends on or changes the state of these two queues jointly. The notions of "vertex" and "queue" will often be used interchangeably; "vertex"

will be preferred when emphasizing topological aspects, and "queue" will be used when referring to queueing processes.

A network of Markovian queues is considered. Each queue has a single server and finite space capacity. The state space associated to a vertex/queue set $\mathcal{W}$ is defined as

$$\mathfrak{N}(\mathcal{W}) \quad = \quad \bigtimes_{i\in\mathcal{W}} \{0,1,\ldots,\ell_i\} \tag{1}$$

where ℓ_i is the space capacity of queue i and $\times$ is the Cartesian product; the resulting set $\mathfrak{N}(\mathcal{W})$ contains all possible state combinations of all queues in $\mathcal{W}$. Denoting by $\mathbf{N} = \mathbf{N}(\tau) \in \mathfrak{N}(\mathcal{V})$ the random vector of all queue states in the network at real-valued time τ, the dynamics of the joint distribution of $\mathbf{N}$ are guided by the following linear system of differential equations (Reibman; 1991):

$$\frac{d}{d\tau}P(\mathbf{N}=\mathbf{y}) \quad = \quad \sum_{\mathbf{x}\in\mathfrak{N}(\mathcal{V})} t^{\mathbf{y}}_{\mathbf{x}} P(\mathbf{N}=\mathbf{x}) \tag{2}$$

where $\frac{d}{d\tau}P$ is the time derivative of P, both $\mathbf{x} = (x_i)$ and $\mathbf{y} = (y_i)$ are elements of $\mathfrak{N}(\mathcal{V})$, and $t^{\mathbf{y}}_{\mathbf{x}}$ is the transition rate from state $\mathbf{x}$ into state $\mathbf{y}$.

The unit of a transition rate is time^{-1}. Conservation of probability mass (the probabilities of being in any possible state must sum up to one) is established by defining the departure rates

$$t^{\mathbf{x}}_{\mathbf{x}} \quad = \quad - \sum_{\mathbf{y}\in\mathfrak{N}(\mathcal{V}),\, \mathbf{y}\neq\mathbf{x}} t^{\mathbf{y}}_{\mathbf{x}}, \tag{3}$$

which captures the effect that state $\mathbf{x}$ leading to state $\mathbf{y}$ reduces the probability of remaining in state $\mathbf{x}$ correspondingly.

Moving from one state to another is associated with the occurrence of an event. For each event, the inter-event times are assumed to be independent exponential random variables with rate parameters that may change over time, i.e. $t^{\mathbf{y}}_{\mathbf{x}} = t^{\mathbf{y}}_{\mathbf{x}}(\tau)$ in (2). All transition rates are exogenous.

The model (2) becomes computationally intractable for non-trivial networks since the dimension of the state space $\mathfrak{N}(\mathcal{V})$ is exponential in the number of queues, cf. (1). This work hence proposes a decomposition technique that approximates the transient queue-length distributions of overlapping subnetworks. These distributions can then be used to approximate properties of the high-dimensional joint distribution $P(\mathbf{N})$.

Definition 1. *Denote by a* subnetwork *S any non-empty set of vertices, and let a* subnetwork decomposition *$\mathscr{S}(G)$ of a given graph G be any choice of subnetworks such that each vertex is contained in either one or two subnetworks. Let $\mathcal{V}(S)$ be the set of vertices contained in subnetwork S.*

The subnetwork neighborhood *of any subnetwork S is defined as*

$$\partial S \quad = \quad \{T \in \mathscr{S}(G) \mid T \neq S,\, \mathcal{V}(T) \cap \mathcal{V}(S) \neq \emptyset\}. \tag{4}$$

The vertex neighborhood *of any vertex set $\mathcal{W} \subset \mathcal{V}$ is defined as*

$$\partial\mathcal{W} \quad = \quad \left(\bigcup_{T\in\mathscr{S}(G):\mathcal{V}(T)\cap\mathcal{W}\neq\emptyset} \mathcal{V}(T) \right) \backslash \mathcal{W}. \tag{5}$$

The vertex neighborhood of a subnetwork $S \in \mathscr{S}(G)$ is hence written as $\partial\mathcal{V}(S)$.

In words: The *subnetwork neighborhood* of a given subnetwork consists of all other subnetworks that have at least one common vertex with the given subnetwork. The *vertex neighborhood* of a given vertex set consists of the vertices of all subnetworks that contain at least one element of the given vertex set.

Definition 2. *A subnetwork decomposition $\mathscr{S}(G)$ is called* triangle-free *if for all $S \in \mathscr{S}(G)$ and $T_1, T_2 \in \partial S$ one has $[\mathcal{V}(T_1) \cap \mathcal{V}(T_2)]\backslash\mathcal{V}(S) = \emptyset$.*

The *triangle-free* definition excludes subnetwork configurations where subnetwork S overlaps with subnetworks T_1 and T_2, and T_1 and T_2 overlap with each other outside of S.

Definition 3. *A network G and a corresponding subnetwork decomposition $\mathscr{S}(G)$ and transition rate matrix t are said to allow for* instantaneous local transitions only *if the following holds for all $\mathbf{x}, \mathbf{y} \in \mathfrak{N}(\mathcal{V})$:*

$$t^{\mathbf{y}}_{\mathbf{x}} \neq 0 \quad \Rightarrow \quad \exists\, S \in \mathscr{S}(G) : t^{\mathbf{y}}_{\mathbf{x}} = t^{\mathbf{y}_{\mathcal{V}(S)},\mathbf{z}}_{\mathbf{x}_{\mathcal{V}(S)},\mathbf{z}} \quad \forall\, \mathbf{z} \in \mathfrak{N}(\mathcal{V}\backslash\mathcal{V}(S)). \tag{6}$$

Algorithm 1 Decomposition of a general queueing network

1. For every event that depends on or affects one or more queues, create one subnetwork containing all of the corresponding vertices.
2. Remove vertices that are not contained in any subnetwork because the corresponding queues do not participate in the network dynamics.
3. Repeat one or several of the following steps until a triangle-free subnetwork decomposition is obtained that allows for instantaneous local transitions only.
   - Discard subnetworks that are fully contained in other subnetworks.
   - If several subnetworks form a triangle, merge two or more of them until the resulting configuration is triangle-free.
   - If a vertex is contained in more than two subnetworks, merge two or more of these subnetworks until the vertex is contained in at most two subnetworks.

Allowing for *instantaneous local transitions only* means that every event (and corresponding network state change) can be inscribed in a subnetwork, in that (i) this change only affects states within that subnetwork and (ii) is independent of states outside of that subnetwork.

The following developments hinge on the availability of a triangle-free subnetwork decomposition that allows for instantaneous local transitions only. For the purpose of devising the SLTM, this decomposition will emerge naturally, as described in Section 3.

Algorithm 1 provides a blueprint for the decomposition of a general network. Step 1 creates a finite set of subnetworks. Step 3 reduces the number of subnetworks by discarding or merging them. The algorithm terminates at the latest when only one subnetwork comprising the full original network is left because this constitutes a valid triangle-free subnetwork decomposition that allows for local transitions only.

Further elaboration on how a concrete instance of Algorithm 1 could look is omitted in the present article because (i) this is not necessary for developing the SLTM and (ii) it would depend on how one wishes to balance computational efficiency (resulting from small subnetworks that approximate the joint distribution of only a few queues) and approximation quality (resulting from large subnetworks that capture the joint distribution of many queues) in a concrete queueing network configuration.

2.1.2 Subnetwork dynamics

Some notational conventions will simplify the presentation. (i) The subset of elements of $\mathbf{x} \in \mathfrak{N}(\mathcal{V})$ that is also contained in the state space $\mathfrak{N}(\mathcal{W})$, $\mathcal{W} \subset \mathcal{V}$, is written as $\mathbf{x}_{\mathcal{W}}$. (ii) A summation of the form $\sum_{\mathbf{z} \in \mathfrak{N}(\mathcal{W})}(\cdot)$ with $\mathcal{W} \subset \mathcal{V}$ is abbreviated as $\sum_{\mathbf{z}}(\cdot)$ and the concrete definition of $\mathbf{z}$ is provided in the context.

In this section, a tractable approximation is derived for the dynamics of any subnetwork $S \in \mathscr{S}(G)$, i.e. of $\frac{d}{d\tau}P(\mathbf{N}_{\mathcal{V}(S)})$. For this, the vectors $\mathbf{x}, \mathbf{y} \in \mathfrak{N}(\mathcal{V})$ in (2) are split into their components representing the states of queues in $\mathcal{V}(S)$, in its neighborhood $\partial\mathcal{V}(S)$, and in the remaining network $\mathcal{V}\backslash[\mathcal{V}(S) \cup \partial\mathcal{V}(S)]$. Specifically, $\mathbf{x} = (\mathbf{m}, \mathbf{r}, \mathbf{v})$ and $\mathbf{y} = (\mathbf{n}, \mathbf{s}, \mathbf{w})$ with $\mathbf{m}, \mathbf{n} \in \mathfrak{N}(\mathcal{V}(S))$ and $\mathbf{r}, \mathbf{s} \in \mathfrak{N}(\partial\mathcal{V}(S))$ and $\mathbf{v}, \mathbf{w} \in \mathfrak{N}(\mathcal{V}\backslash[\mathcal{V}(S) \cup \partial\mathcal{V}(S)])$. Substituting this in (2) and summing both sides of this equation over all $(\mathbf{s}, \mathbf{w}) \in \mathfrak{N}(\mathcal{V}\backslash\mathcal{V}(S))$ yields

$$\frac{d}{d\tau}P(\mathbf{N}_{\mathcal{V}(S)} = \mathbf{n}) \quad = \quad \sum_{\mathbf{s},\mathbf{w}} \sum_{\mathbf{m},\mathbf{r},\mathbf{v}} t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{m},\mathbf{r},\mathbf{v}} P(\mathbf{N} = (\mathbf{m}, \mathbf{r}, \mathbf{v})). \tag{7}$$

To guide the eye, summations over multiple arguments are here and in the following split into (at least) one sum over all final states and one sum over all initial states of a considered transition. Using $P(\mathbf{N} = (\mathbf{m}, \mathbf{r}, \mathbf{v})) = P(\mathbf{N}_{\mathcal{V}\backslash\mathcal{V}(S)} = (\mathbf{r}, \mathbf{v}) \mid \mathbf{N}_{\mathcal{V}(S)} = \mathbf{m})P(\mathbf{N}_{\mathcal{V}(S)} = \mathbf{m})$, (7) is rearranged into

$$\frac{d}{d\tau}P(\mathbf{N}_{\mathcal{V}(S)} = \mathbf{n}) \quad = \quad \sum_{\mathbf{m}} \left[\sum_{\mathbf{s},\mathbf{w}} \sum_{\mathbf{r},\mathbf{v}} t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{m},\mathbf{r},\mathbf{v}} P(\mathbf{N}_{\mathcal{V}\backslash\mathcal{V}(S)} = (\mathbf{r}, \mathbf{v}) \mid \mathbf{N}_{\mathcal{V}(S)} = \mathbf{m}) \right] P(\mathbf{N}_{\mathcal{V}(S)} = \mathbf{m}), \tag{8}$$

where the term in square brackets functions like a state-dependent transition rate from subnetwork state $\mathbf{m}$ to subnetwork state $\mathbf{n}$. This – so far exact – expression is the basis for the proposed queueing network decomposition model.

Definition 4. *For a given $(G, \mathscr{S}(G), t)$ that allow for instantaneous local transitions only, the* local transition rates *of any vertex set $\mathcal{W} \subset \mathcal{V}$ are defined as follows, assuming $(\mathbf{m}, \mathbf{r}), (\mathbf{n}, \mathbf{s}) \in \mathfrak{N}(\mathcal{W}) \times \mathfrak{N}(\partial\mathcal{W})$:*

$$t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{W}) \quad = \quad \begin{cases} t^{\mathbf{n},\mathbf{s},\mathbf{v}}_{\mathbf{m},\mathbf{r},\mathbf{v}} & \text{if } \mathbf{m} \neq \mathbf{n} \text{ with } \mathbf{v} \in \mathfrak{N}(\mathcal{V}\backslash(\mathcal{W} \cup \partial\mathcal{W})) \text{ arbitrary} \\ -\sum_{\mathbf{a} \in \mathfrak{N}(\mathcal{W}), \mathbf{a} \neq \mathbf{m}} \sum_{\mathbf{b} \in \mathfrak{N}(\partial\mathcal{W})} t^{\mathbf{a},\mathbf{b}}_{\mathbf{m},\mathbf{r}}(\mathcal{W}) & \text{if } (\mathbf{m}, \mathbf{r}) = (\mathbf{n}, \mathbf{s}) \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The first row of (9) expresses state transitions that involve queues in $\mathcal{W}$ independently of the states of queues that are neither in $\mathcal{W}$ nor its neighborhood $\partial\mathcal{W}$. This is feasible because (i) according to (5), $\mathcal{W} \cup \partial\mathcal{W}$ comprises the queues of all subnetworks into which events in $\mathcal{W}$ could possibly be inscribed and (ii) Definition 3 ensures that the states of queues in $\mathcal{V}\backslash(\mathcal{W} \cup \partial\mathcal{W})$ do not affect events in $\mathcal{W}$. The second row of (9) ensures a proper transition rate matrix specific to $\mathcal{W}$, in that it defines its main diagonal elements as a function of the rates of departure from the corresponding states, cf. (3). The third row excludes from consideration all events that do not affect queues in $\mathcal{W}$.

Proposition 1. *Let $(G, \mathscr{S}(G), t)$ allow for instantaneous local transitions only. Let $\mathcal{W} \subset \mathcal{V}$ and $(\mathbf{m}, \mathbf{r}), (\mathbf{n}, \mathbf{s}) \in \mathfrak{N}(\mathcal{W}) \times \mathfrak{N}(\partial\mathcal{W})$. Then, the time derivative of the state distribution of $\mathcal{W}$ can be expressed as a function of only the distribution of $\mathcal{W}$, $\partial\mathcal{W}$ and of the corresponding local transition rates* (9)*:*

$$\frac{d}{d\tau} P(\mathbf{N}_{\mathcal{W}} = \mathbf{n}) \quad = \quad \sum_{\mathbf{s}} \sum_{\mathbf{m},\mathbf{r}} t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{W}) P(\mathbf{N}_{\mathcal{W},\partial\mathcal{W}} = (\mathbf{m}, \mathbf{r})). \tag{10}$$

Proof. See Appendix A.1. ■

This means that one can compute the instantaneous temporal change of the state distribution of a queue set $\mathcal{W}$ by only looking at these queues and their neighbors in $\partial\mathcal{W}$, without considering the state of any other queue in the network.

This last, exact result is now taken as the starting point for devising a decomposition scheme where the joint queue dynamics of a full network are approximated through many overlapping subnetworks. This requires a formulation where the state of a given subnetwork can be updated without having to condition on the full network state – otherwise, one would be back solving the full model (8). Definition 4 delivers half the solution to this problem because it yields the local (i.e. not network-wide) transition rates needed to define the exact subnetwork dynamics in Proposition 1. However, this proposition also uses the joint distribution of all queues in the considered subnetwork and its neighborhood. This joint distribution is in the present decomposition scheme not exactly represented but needs to be approximately recovered from the involved subnetwork distributions.

Letting $S \in \mathscr{S}(G)$, with $\mathscr{S}(G)$ being triangle-free, the central approximation of the proposed model consists of the following two steps:

$$P(\mathbf{N}_{\partial\mathcal{V}(S)} \mid \mathbf{N}_{\mathcal{V}(S)}) \quad \approx \quad \prod_{T \in \partial S} P(\mathbf{N}_{\mathcal{V}(T)\backslash\mathcal{V}(S)} \mid \mathbf{N}_{\mathcal{V}(S)}) \tag{11}$$

$$\approx \quad \prod_{T \in \partial S} P(\mathbf{N}_{\mathcal{V}(T)\backslash\mathcal{V}(S)} \mid \mathbf{N}_{\mathcal{V}(T)\cap\mathcal{V}(S)}). \tag{12}$$

The first expression (11) approximates the conditional distribution $P(\mathbf{N}_{\partial\mathcal{V}(S)} \mid \mathbf{N}_{\mathcal{V}(S)})$ of queue states adjacent to subnetwork S given queue states within S through a factorization over all subnetworks T in the neighborhood of S. Since the subnetworks entering this product have by Definition 2 no mutual overlap, this expression can be interpreted as the exact consequence of assuming for all $T \in \partial S$ conditional independence between their respective $\mathbf{N}_{\mathcal{V}(T)\backslash\mathcal{V}(S)}$ given $\mathbf{N}_{\mathcal{V}(S)}$. The proposed model is Markovian along the time line, but this does not imply that the resulting joint state distributions are Markovian along paths in the network, as is illustrated with a simple example immediately below in Section 2.1.3.

The second approximation (12) then considers $\mathbf{N}_{\mathcal{V}(T)\backslash\mathcal{V}(S)}$ to be independent of the states in S that are not in T conditional on the states that are in S and in T. The subsequent Section 2.1.3 also illustrates that this is not an inherent model property but an approximation. The resulting formula (12) is operational because each of its factors can be computed from the state distribution $P(\mathbf{N}_{\mathcal{V}(T)})$ of the corresponding subnetwork T alone.

The proposed network model can now be stated. It assumes a triangle-free subnetwork decomposition $\mathscr{S}(G)$ to be given that allows for instantaneous local transitions only. The model defines an approximate distribution

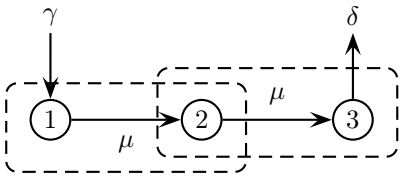


Figure 1: Tandem network

$\Phi_S(\mathbf{N}_{\mathcal{V}(S)})$ of the stochastic state vector $\mathbf{N}_{\mathcal{V}(S)}$ of every subnetwork $S \in \mathscr{S}(G)$. It combines the exact local dynamics (10) with the approximation (12). Letting $(\mathbf{m}, \mathbf{r}), (\mathbf{n}, \mathbf{s}) \in \mathfrak{N}(\mathcal{V}(S)) \times \mathfrak{N}(\partial\mathcal{V}(S))$, it reads as follows:

$$\frac{d}{d\tau}\Phi_S(\mathbf{N}_{\mathcal{V}(S)} = \mathbf{n}) = \sum_{\mathbf{m}} \left[\sum_{\mathbf{r},\mathbf{s}} t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{V}(S)) \Psi_S(\mathbf{N}_{\partial\mathcal{V}(S)} = \mathbf{r} \mid \mathbf{N}_{\mathcal{V}(S)} = \mathbf{m}) \right] \Phi_S(\mathbf{N}_{\mathcal{V}(S)} = \mathbf{m}) \quad (13)$$

$$\Psi_S(\mathbf{N}_{\partial\mathcal{V}(S)} = \mathbf{r} \mid \mathbf{N}_{\mathcal{V}(S)} = \mathbf{m}) = \prod_{T \in \partial S} \Phi_T(\mathbf{N}_{\mathcal{V}(T)\setminus\mathcal{V}(S)} = \mathbf{r}_{\mathcal{V}(T)\setminus\mathcal{V}(S)} \mid \mathbf{N}_{\mathcal{V}(T)\cap\mathcal{V}(S)} = \mathbf{m}_{\mathcal{V}(T)\cap\mathcal{V}(S)}). \quad (14)$$

Equation (13) is the approximation model's counterpart of the exact model (8), with Φ_S being an approximation of the exact queue state distribution of subnetwork S. Differently from (8), the term in square brackets now only involves local transition rates and an approximation Ψ_S of the states of subnetworks in the neighborhood of S given the state of S. The definition of Ψ is given in (14). It makes the same approximations as (12), only that its right-hand side evaluates approximate subnetwork distributions Φ.

Proposition 2. *Let $(G, \mathscr{S}(G), t)$ be triangle-free and allow for instantaneous local transitions only. Consider the two subnetworks $S, T \in \mathscr{S}(G)$ with $S \neq T$ and $\mathcal{W} = \mathcal{V}(S) \cap \mathcal{V}(T) \neq \emptyset$. Let Φ_S and Φ_T be probability distributions over $\mathfrak{N}(\mathcal{V}(S))$ and $\mathfrak{N}(\mathcal{V}(T))$, respectively. Then, the model* (13),(14) *has the following property:*

$$\Phi_S(\mathbf{N}_{\mathcal{W}}) = \Phi_T(\mathbf{N}_{\mathcal{W}}) \quad \Rightarrow \quad \frac{d}{d\tau}\Phi_S(\mathbf{N}_{\mathcal{W}}) = \frac{d}{d\tau}\Phi_T(\mathbf{N}_{\mathcal{W}}) \quad (15)$$

where $\mathbf{N}_{\mathcal{W}} \in \mathfrak{N}(\mathcal{W})$. That is, if any two subnetwork distributions have identical marginals for their common set of queues at some point in time, the marginals will remain identical at all other points in time.

Proof. See Appendix A.2. ■

Proposition 2 states a key feature of the proposed decomposition approach: The model (13), (14) maintains mutually consistent overlapping subnetwork distributions, without any need to introduce supplementary distributional adjustments or constraints.

2.1.3 Illustration of the adopted approximations

The sole purpose of this section is to illustrate the approximations made in (11) and (12); no additional modeling concepts are introduced.

The queueing network displayed in Figure 1 is considered. It consists of three queues 1, 2, 3 in tandem, with each queue having a (for simplicity unitless) flow capacity of $\mu = 1$ and a space capacity of $\ell = 1$. Jobs arrive to queue 1 at rate $\gamma = 1$ and leave from queue 3 at rate $\delta = 1$. The transition rate matrix between the eight binary states of this system is displayed in Table 1. The dashed lines in Figure 1 circumscribe two subnetworks; this decomposition is triangle-free (Def. 2) and allows for local transitions only (Def. 3).

The stationary state of this system is subsequently analyzed; this keeps the presentation simple yet suffices to clarify the points of interest. Denoting Table 1's transition rate matrix by T and stacking the stationary probability of every network state into a column vector $\boldsymbol{\pi}$, the stationary state distribution is defined by the system

$$T^T\boldsymbol{\pi} = \mathbf{0} \quad (16)$$

$$\mathbf{1}^T\boldsymbol{\pi} = 1 \quad (17)$$

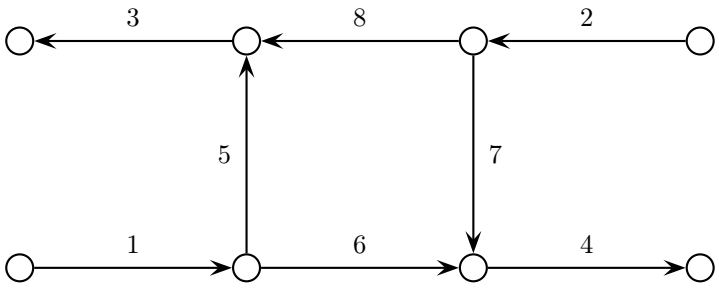
where $\mathbf{0}$ and $\mathbf{1}$ are all-zero resp. all-one column vectors of suitable dimension and superscript T denotes the transpose. Solving this system yields the state probabilities $\pi(n_1, n_2, n_3)$ displayed in the fourth column of Table 2. The remaining columns of this table are all derived from these values by summing out dimensions and/or conditioning.

Table 1: Transition rates in tandem network

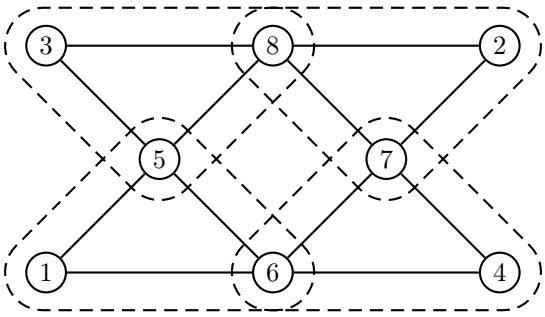
	to	n_1	0	0	0	0	1	1	1	1
from		n_2	0	0	1	1	0	0	1	1
n_1	n_2	n_3	0	1	0	1	0	1	0	1
0	0	0	-1				$\gamma = 1$			
0	0	1	$\delta = 1$	-2				$\gamma = 1$		
0	1	0		$\mu = 1$	-2				$\gamma = 1$	
0	1	1			$\delta = 1$	-2				$\gamma = 1$
1	0	0			$\mu = 1$		-1			
1	0	1				$\mu = 1$	$\delta = 1$	-2		
1	1	0						$\mu = 1$	-1	
1	1	1							$\delta = 1$	-1

Table 2: Stationary state distribution and derived quantitites

n_1	n_2	n_3	$\pi(n_1, n_2, n_3)$	$\pi(n_1, n_3 \mid n_2)$	$\pi(n_1 \mid n_2) \cdot \pi(n_3 \mid n_2)$	$\pi(n_3 \mid n_1, n_2)$	$\pi(n_3 \mid n_2)$
0	0	0	0.0714	0.1428	0.1632	0.5	0.5714
0	0	1	0.0714	0.1428	0.1224	0.5	0.4286
0	1	0	0.1429	0.2858	0.3062	0.6668	0.7144
0	1	1	0.0714	0.1428	0.1224	0.3332	0.2856
1	0	0	0.2143	0.4286	0.4082	0.5999	0.5714
1	0	1	0.1429	0.2858	0.3062	0.4001	0.4286
1	1	0	0.2143	0.4286	0.4082	0.7501	0.7144
1	1	1	0.0714	0.1428	0.1632	0.2499	0.2856



(a) Road network



(b) Queueing network with subnetworks

Figure 2: Decomposition of example network

The fifth and sixth column compare the exact joint distribution $\pi(n_1, n_3 \mid n_2)$ of queues 1 and 3 given queue 2 to an expression $\pi(n_1 \mid n_2) \cdot \pi(n_3 \mid n_2)$ that would be equivalent if the outer queues 1 and 3 were conditionally independent given the middle queue 2. The table reveals that this is not the case, illustrating that the conditioning of adjacent subnetworks on a given intermediate subnetwork in (11) is an approximation.

Similarly, the two last columns compare the full conditional distribution $\pi(n_3 \mid n_1, n_2)$ of queue 3 on all other queues to the result of conditioning it only on its adjacent queue 2, i.e. to $\pi(n_3 \mid n_2)$. Different numbers are obtained, illustrating that the incomplete conditioning in (12) is an approximation.

2.2 Model validation

The purpose of the experiments presented here is to investigate the capability of the proposed approximation model (13), (14) to capture uni- and multivariate queue state distributions in a network with intricate dynamics. The analytical approximations are compared to estimates obtained from an event-based queueing network simulator that generates realizations of network state trajectories according to the exact model (2). Statistics are computed from 10^7 replications of the simulation.

A road traffic scenario is considered, using the road network shown in Figure 2(a) where vertices represent intersections and edges represent road segments. This network could describe an arterial consisting of one westbound road (road segments 2,8,3) and one eastbound road (segments 1,6,4), between which U-turns are enabled by road segments 5 and 7. All roads are directed (as indicated by the arrows) and have a single lane. Following Osorio (2010, Chap. 4), this road network is now modeled through a queueing network by (i) representing each link by a single server queue with finite space capacity ℓ, independent and exponentially distributed service times, external network arrivals that constitute a Poisson process, and (ii) representing each possible turning move in every road intersection by a corresponding edge in the queueing network. The resulting queueing network becomes the line graph of the road network (Balakrishnan; 1997). It is shown in Figure 2(b). The circles represent queues. Two queues are connected by a solid line if there exists a network state transition that depends on both queues or affects both queues. These state transitions and the subnetwork decomposition (dashed) are detailed further below.

This queueing representation of a road network leads to a simplistic representation of real road traffic dynamics because it only captures delay caused by congestion but neglects the finite speed at which traffic states at

Table 3: Transition rates in test network

description	**final state n**	**rate $t^{\mathbf{n}}_{\mathbf{m}}$**	**condition**
arrival to 1	$m_1 + 1$	γ_1	$m_1 < \ell_1$
arrival to 2	$m_2 + 1$	γ_2	$m_2 < \ell_2$
departure from 3	$m_3 - 1$	μ_3	$m_3 > 0$
departure from 4	$m_4 - 1$	μ_4	$m_4 > 0$
transition from 1 to 5	$m_1 - 1, m_5 + 1$	$p_{15}\mu_1$	$m_1 > 0, m_5 < \ell_5, m_6 < \ell_6$
transition from 1 to 6	$m_1 - 1, m_6 + 1$	$p_{16}\mu_1$	$m_1 > 0, m_5 < \ell_5, m_6 < \ell_6$
transition from 2 to 7	$m_2 - 1, m_7 + 1$	$p_{27}\mu_2$	$m_2 > 0, m_7 < \ell_7, m_8 < \ell_8$
transition from 2 to 8	$m_2 - 1, m_8 + 1$	$p_{28}\mu_2$	$m_2 > 0, m_7 < \ell_7, m_8 < \ell_8$
transition from 5 to 3	$m_5 - 1, m_3 + 1$	μ_5	$m_5 > 0, m_3 < \ell_3, m_8 = 0$
transition from 8 to 3	$m_8 - 1, m_3 + 1$	μ_8	$m_8 > 0, m_3 < \ell_3$
transition from 7 to 4	$m_7 - 1, m_4 + 1$	μ_7	$m_7 > 0, m_4 < \ell_4, m_6 = 0$
transition from 6 to 4	$m_6 - 1, m_4 + 1$	μ_6	$m_6 > 0, m_4 < \ell_4$

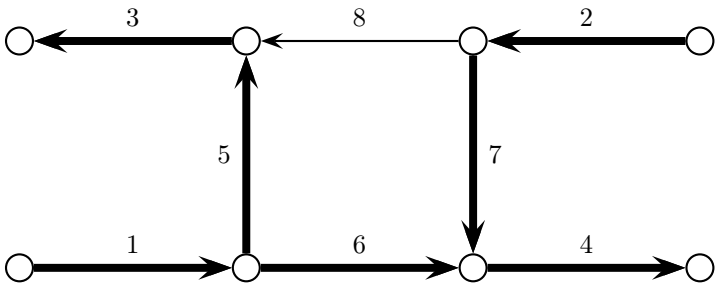
different coordinates propagate (in the form of kinematic waves) along the link. These deficiencies will be removed in the SLTM *road* network model presented in Section 3. The present case study merely aims at illustrating the previously developed *queueing* network model.

The non-zero and non-diagonal transition rates of this system are given in Table 3. The first column describes the different possible events. The second column indicates those elements of the state vector that have changed after the corresponding event, assuming an initial state $\mathbf{m} = (m_i)$. The third column gives the transition rate, and the fourth column indicates the condition under which the transition is feasible. The symbols γ, μ, and ℓ represent exogenous arrival rates, queue service rates, and queue space capacities, respectively. In addition, p_{ij} represents the transition probability from upstream queue i into downstream queue j. The diagonal transition rates, i.e. the rates of departure from each state, can be obtained through (3), as explained in Section 2.1.1.

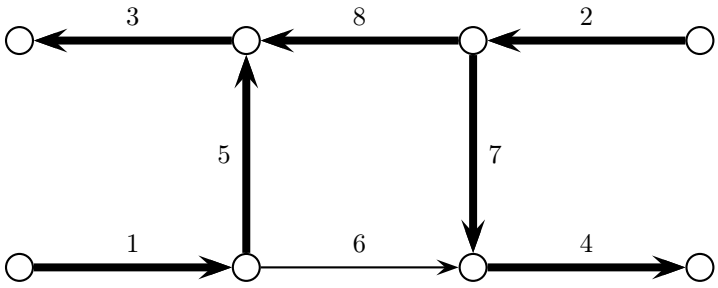
Vehicles enter the network by joining queue 1 or 2, and they leave the network through queue 3 or 4. From queue 1, they can either go straight into queue 6 or initiate a U-turn by entering queue 5. Either turn is only allowed if both downstream queues are non-full. This mimics spillback effects in road networks, where vehicles attempting to enter a full road block the traffic on the intersection upstream of that road. Vehicles continuing straight into queue 6 leave the network through queue 4. U-turning vehicles leave the network through queue 3. A transition from queue 5 to queue 3 is only allowed if there is no vehicle in queue 8. This mimics a prioritized road intersection where the merging traffic (from queue 5) yields to the through traffic (from queue 8). A symmetric logic applies to vehicles entering through queue 2.

The concrete parameters used are as follows, with all rates and flow capacities being given in vehicles per second. The space capacity of all queues is $\ell_i = 10$ vehicles, $i = 1 \dots 8$. Vehicles enter the network (with losses, meaning that vehicles that cannot enter due to spillback are discarded) at a rate of $\gamma_1 = \gamma_2 = 1.25$ into queue 1 and 2. They continue straight with probability $p_{16} = p_{28} = 2/3$ and perform a U-turn with probability $p_{15} = p_{27} = 1/3$. The queue service rates within the network are $\mu_1 = \mu_2 = \mu_5 = \mu_6 = \mu_7 = \mu_8 = 10$, which is on average sufficient to serve the demand. The outgoing queues 3 and 4, however, constitute bottlenecks with a low service rate of $\mu_3 = \mu_4 = 1$ each. Since the overall demand ($\gamma_1 + \gamma_2 = 2.5$) exceeds the overall network exit flow capacity ($\mu_3 + \mu_4 = 2$), congestion arises at the exit bottlenecks and spreads throughout the network.

The symmetric configuration of this network leads to complex congestion patterns. This can be clarified by analyzing the network first under the assumption that all queues are deterministic. In this setting, the service time of queue i would no longer be exponentially distributed but be deterministic and equal to $1/\mu_i$. Under this assumption, the network has two stable stationary congestion patterns, which are shown in Figure 3. In the first case, there is an unhindered flow from queue 1 through queue 6 to queue 4. Because of this, departures from queues 7 are held back, which in turn blocks queue 2. In consequence, there also is no straight flow from queue 2 through queue 8 to queue 3, meaning that U-turns from queue 1 through queue 5 into queue 3 are unhindered. The second case is symmetric to the first one, only that queue 2 sends unhindered and queue 1 is held back. Returning to the stochastic perspective (with exponentially distributed service times), one hence can expect a symmetric, bi-modal distribution of network states.



(a) Unhindered inflow from queue 1; queue 2 is blocked back.



(b) Unhindered inflow from queue 2; queue 1 is blocked back.

Figure 3: Stable stationary congestion patterns

Given Table 3, the stochastic traffic flow dynamics on this network can be evaluated using (7). In order to tackle the exponential complexity of this network model, the queueing network is decomposed into the four subnetworks indicated by dashed lines in Figure 2(b). These subnetworks are subsequently labeled according to the queues they comprise as 156, 278, 358, and 467. Inspecting the overlap of the dashed subnetworks in Fig. 2(b) reveals that this subnetwork decomposition is *triangle-free* (Definition 2). Noting further that the queues referred to in every single row of Table 3 can be inscribed in a single subnetwork leads to the observation that this configuration allows for *instantaneous local transitions only* (Definition 3). All necessary prerequisites to deploy the subnetwork decomposition model (13), (14) are hence satisfied.

The details of this subnetwork decomposition, in particular the evaluation of the *local transition rates* specified in Definition 4, are omitted here to avoid redundancies with Section 3.1, which provides this information when defining the full SLTM.

The initially empty system is simulated for 250 seconds. Figure 4 shows the mean values (column 1) and standard deviations (column 2) of the number of vehicles in each queue over time. Due to the symmetry of the experiment, each row corresponds to two queues. The following observations can be made.

- The analytical model captures very well the transient dynamics of the system, both in terms of queue-length expectations and standard deviations.
- The analytical model also approximates with good precision the stationary expected queue-lengths and their standard deviations.

Proposition 2 ensures the mutual consistency of subnetwork distributions at overlapping queues, but it does not ensure their network-wide consistency in terms of an underlying full joint distribution; neither does it provide a recipe for approximating the joint distribution of two queues that are not elements of the same subnetwork. An approximation scheme is subsequently used, where the joint distribution of two queues is approximated by (i) identifying a sequence of overlapping subnetworks with the first (last) subnetwork in this sequence containing the first (second) queue of interest and then (ii) summing out the states of all other queues contained in this subnetwork sequence. This computation uses only the instantaneous subnetwork distributions Φ, which are readily available from solving the system of differential equations (13), (14) forwards through time.

Figure 2b reveals a circular arrangement of the subnetworks, meaning that for each pair of queues there are two sequences of subnetworks one can traverse to connect them: one clockwise, and one counter-clockwise.

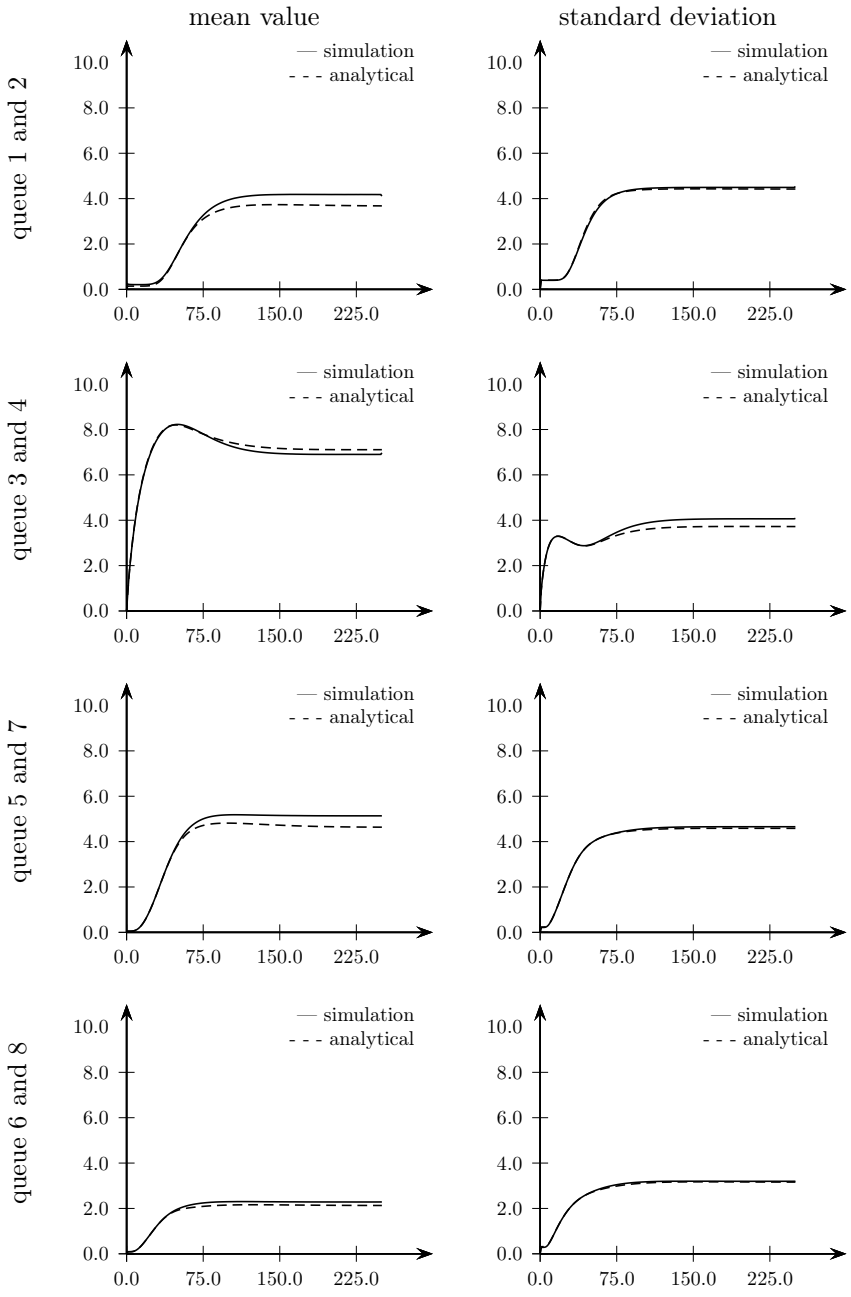


Figure 4: Queue-length expectations and standard deviations [vehicles] over time [s]

Consider, for example, the two-dimensional joint distribution of queue 1 and 2. In counter-clockwise direction, this joint is approximated by considering subnetworks 156, 467, and 278 only:

$$\begin{aligned} & P_{1(67)2}(X_1 = x_1, X_2 = x_2) \\ = & \sum_{x_6} \sum_{x_7} \Phi_{156}(X_1 = x_1 \mid X_6 = x_6) \Phi_{467}(X_6 = x_6 \mid X_7 = x_7) \Phi_{278}(X_2 = x_2, X_7 = x_7) \end{aligned} \tag{18}$$

where the subscript 1(67)2 indicates that the joint of 1 and 2 is computed by summing out states along the path 67. Similarly, the computation in clockwise direction through subnetworks 156, 358, 278 yields

$$\begin{aligned} & P_{1(58)2}(X_1 = x_1, X_2 = x_2) \\ = & \sum_{x_5} \sum_{x_8} \Phi_{156}(X_1 = x_1 \mid X_5 = x_5) \Phi_{358}(X_5 = x_5 \mid X_8 = x_8) \Phi_{278}(X_2 = x_2, X_8 = x_8). \end{aligned} \tag{19}$$

In the following, the analytical approximation of any two-dimensional joint distribution is computed along the shorter of the two possible paths. The symmetry of the considered example ensures that for all queue pairs that are connected by two paths of equal length the joint distributions along both paths are identical.

Figures 5 through 8 show all two-dimensional stationary joint distributions of the given system. The first column visualizes the bivariate joint estimated via simulation. The second column shows the corresponding analytical approximation. Every row shows the joint distribution for one or two pairs of queues, where the queue indices of the pair(s) are given within parenthesis. The state of the first queue in each pair is plotted along the x-axis, and the state of the second queue is plotted along the y-axis. When two queue pairs are indicated in a row, these two pairs have an identical joint distribution because of the experiment's symmetry.

The simulation-based joint distributions, which constitute the ground truth to be approximated by the analytical model, are given some interpretation first. All of these distributions are multi-modal, with most of their probability mass concentrated at extreme state configurations where at least one queue is either empty or full. This corresponds well to the intuition of a system that oscillates between the two congestion patterns given in Figure 3. Indeed, most probability peaks match one of these patterns, with the remaining probability mass being distributed along states that correspond to transitions between these patterns. An example configuration is selected to clarify this.

Consider the last row (queues 1 and 5) in Figure 5. In congestion pattern (a) of Figure 3, both queues carry unhindered flow and hence low occupancy, corresponding to the probability peak at coordinates (0, 0). In pattern (b), congestion spills back across both queues, resulting in high occupancies and the corresponding probability peak at coordinates (10, 10). The remaining probability mass is distributed over states that are visited when transitioning between these extremes. A related phenomenon can be found in the second row (queues 1 and 7) of Figure 6: now, congestion pattern (a) implies low occupancy on queue 1 and high occupancy on queue 7 and a corresponding probability peak at coordinates (0, 10), whereas congestion pattern (b) leads to high occupancy on queue 1, low occupancy on queue 7, and a probability peak at coordinates (10, 0). The symmetric and opposite behavior of queues 5 and 7 in these two examples matches the second row of Figure 8: Congestion pattern (a) implies that queue 5 is almost always uncongested and queue 7 is almost always congested, while pattern (b) implies the opposite.

Comparing now the analytical model predictions to their simulation-based counterparts, the following qualitative observations can be made.

- The analytical model captures very well the absence of probability mass in the center of all histograms.
- The analytical model reproduces the probability peak patterns with overall good precision. However, some under-estimations (e.g. for $P(X_3 = 10, X_4 = 0)$ and $P(X_3 = 0, X_4 = 10)$ in Figure 6) and over-estimations (e.g. for $P(X_5 = 0, X_7 = 0)$ and $P(X_5 = 10, x_7 = 10)$ in Figure 8) remain.

A quantitative perspective on this comparison is adopted in Table 4, which gives summary statistics computed from the distributions of Figures 5-8. The first column indicates the considered queue pairs. The second column states how many queues separate the elements of each pair along their computation path. The third column shows the Kullback-Leibler divergence (Kullback and Leibler; 1951) between simulated distribution and analytical approximation, which is computed as follows:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{i=0}^{\ell_i} \log_2 \left(\frac{P(i)}{Q(i)} \right) P(i) \tag{20}$$

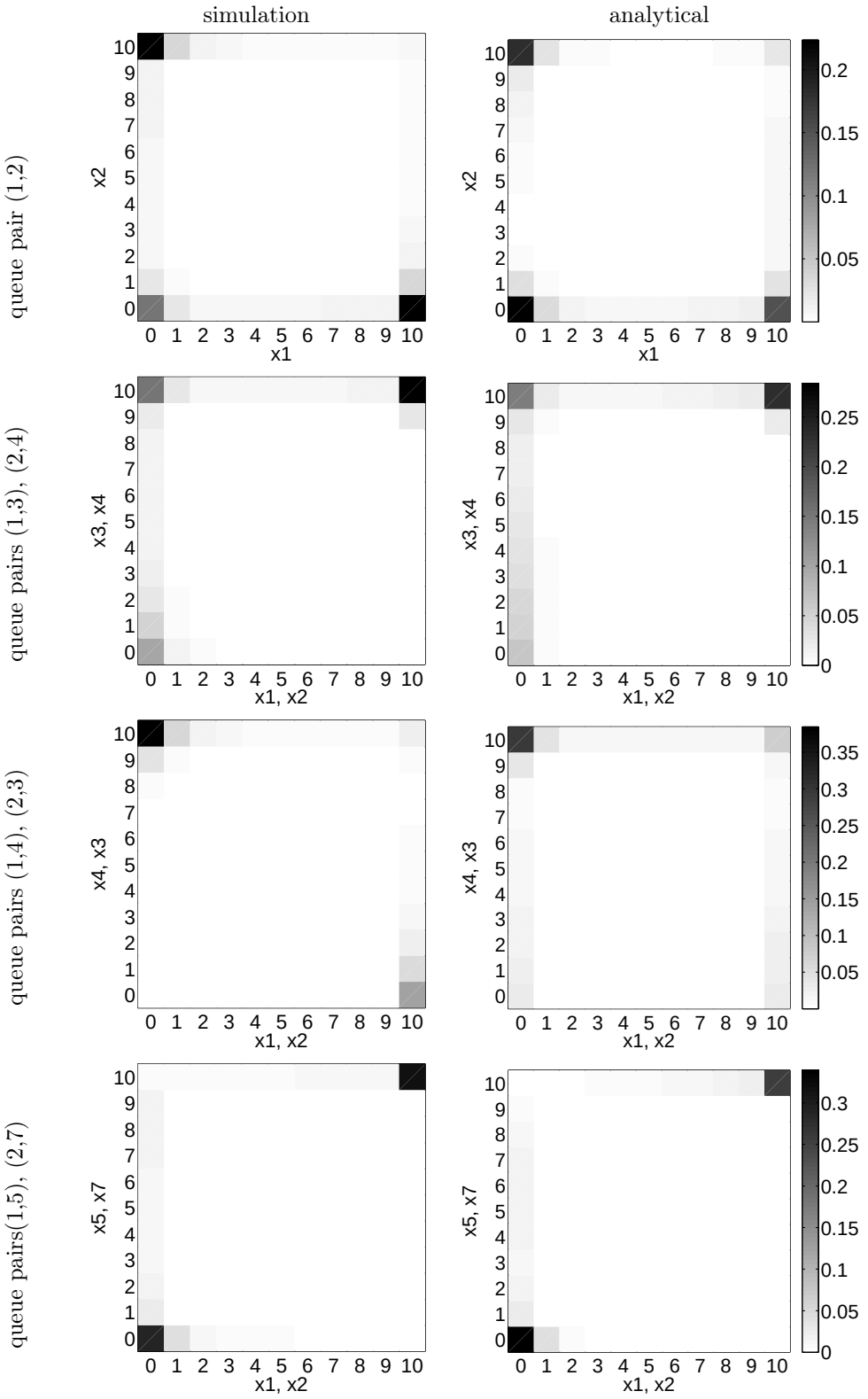


Figure 5: Bivariate queue-length distributions

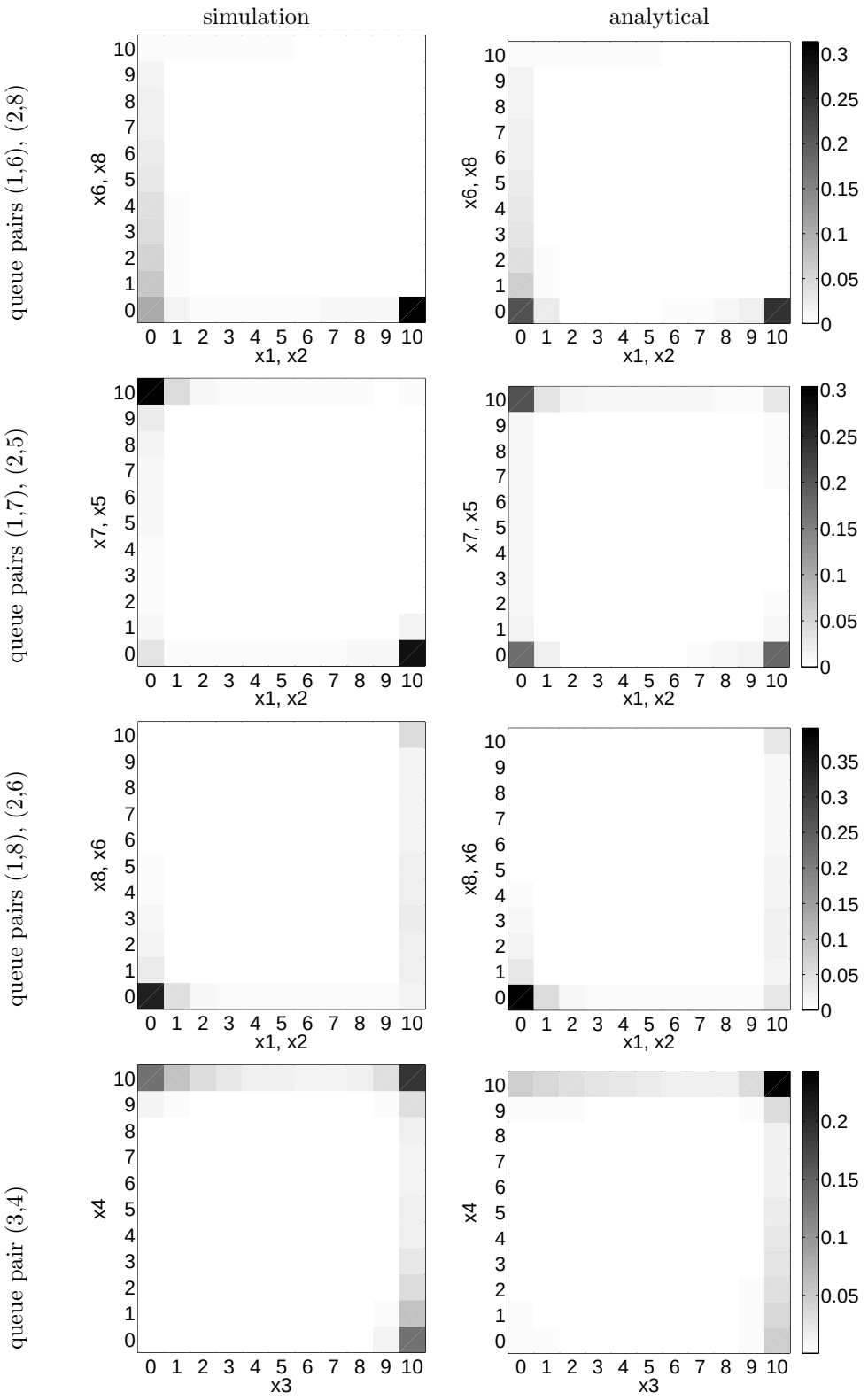


Figure 6: Bivariate queue-length distributions

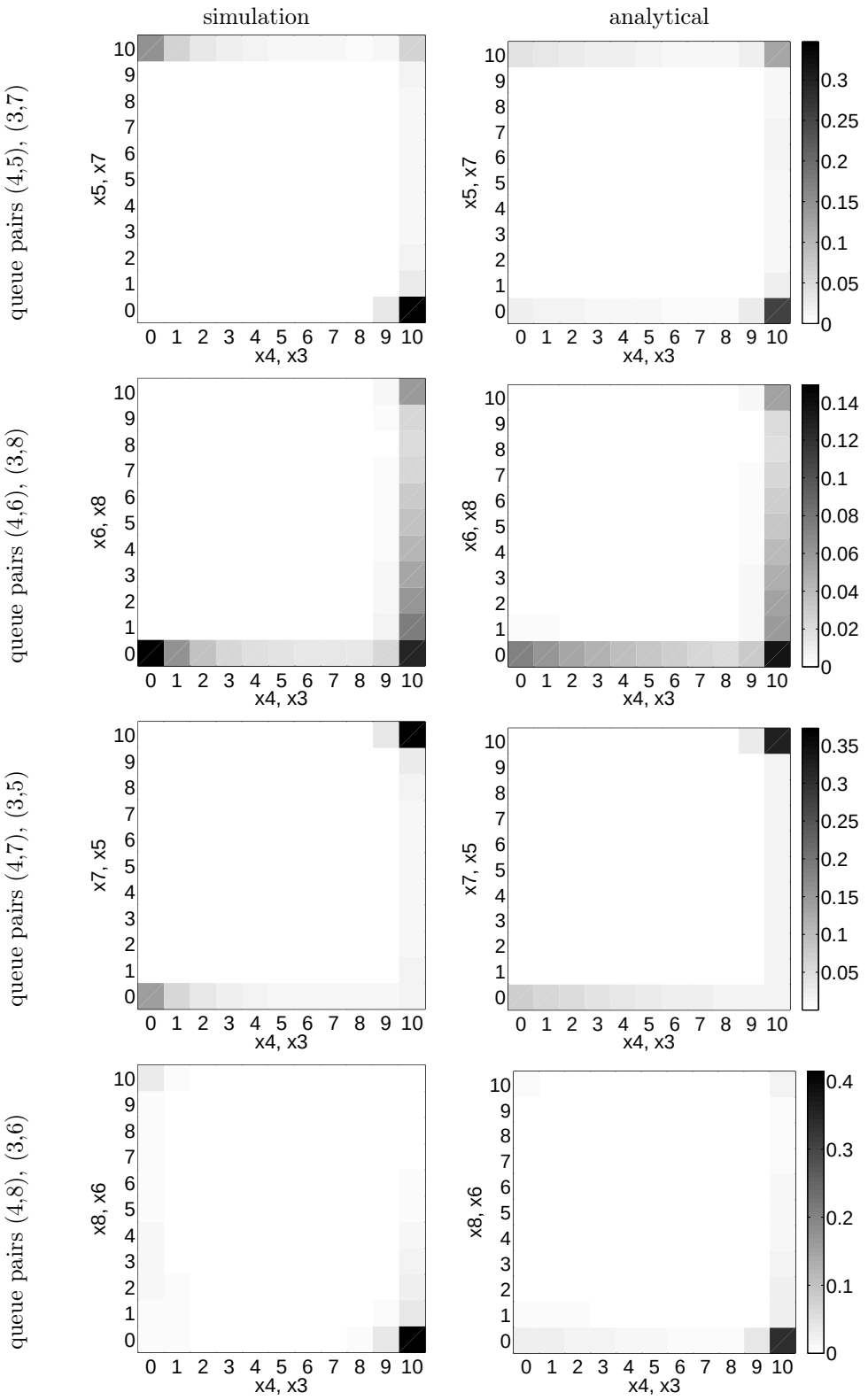


Figure 7: Bivariate queue-length distributions

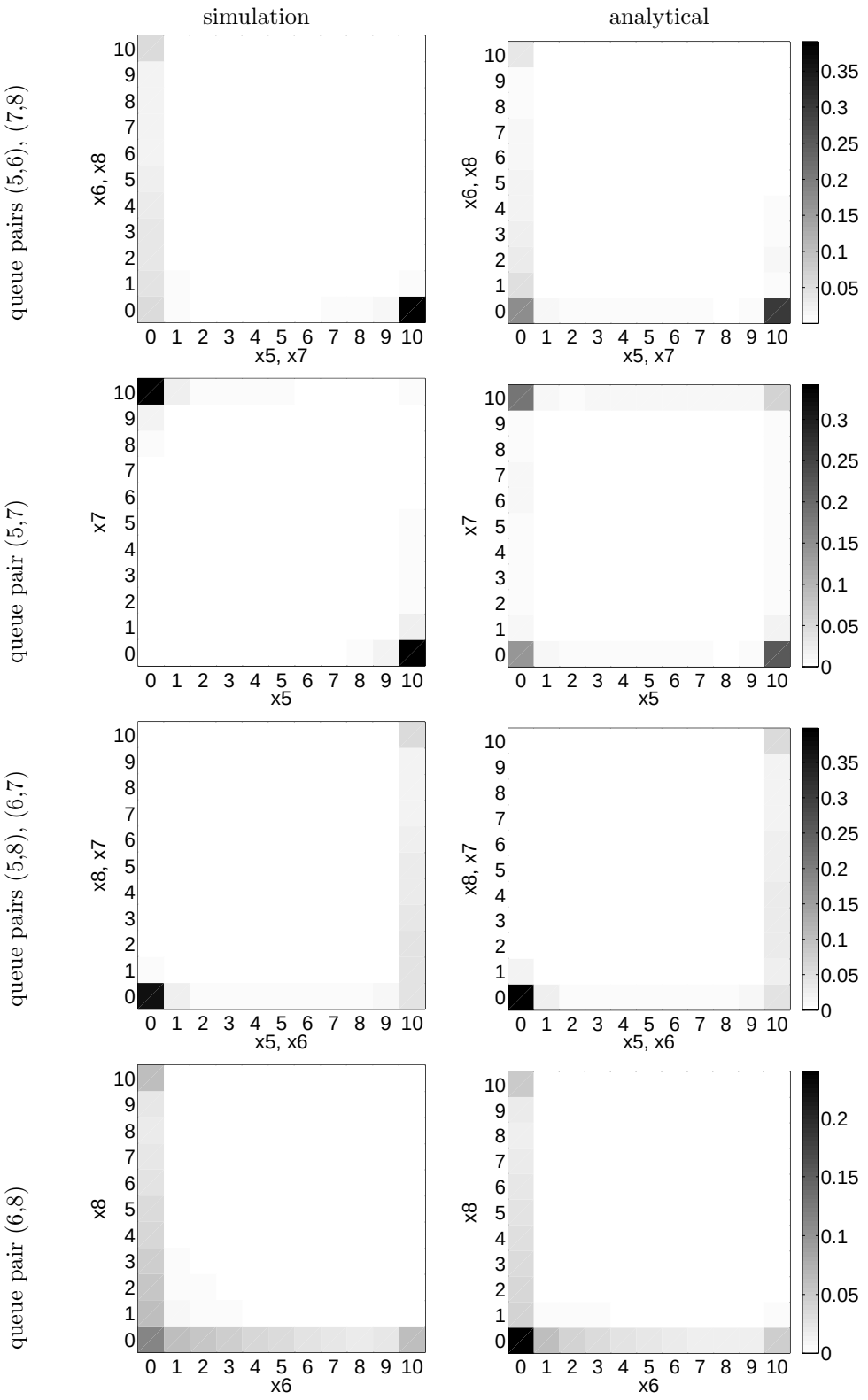


Figure 8: Bivariate queue-length distributions

Table 4: Summary statistics of bivariate joint approximation

queues	**distance**	$D_{\mathrm{KL}}(P \parallel Q)$	$D_{\mathrm{LK}}(P \parallel \text{marginals})$	$D_{\mathrm{LK}}(P \parallel \text{uniform})$
(1,5), (2,7)	0	0.055355	0.717129	3.422993
(1,6), (2,8)	0	0.125268	0.507955	2.796826
(4,6), (3,8)	0	0.095745	0.444516	2.674944
(4,7), (3,5)	0	0.091756	0.735241	3.383012
(5,6), (7,8)	0	0.183747	0.483327	2.839028
(5,8), (6,7)	0	0.036427	0.540676	2.898564
average	0	0.0980	0.5715	3.0026
(1,3), (2,4)	1	0.116154	0.275907	2.855589
(1,4), (2,3)	1	0.418250	0.583807	3.164398
(1,7), (2,5)	1	0.284186	0.503341	3.209555
(1,8), (2,6)	1	0.072494	0.381984	2.673042
(4,5), (3,7)	1	0.410486	0.568681	3.216102
(4,8), (3,6)	1	0.258929	0.317868	2.550483
(5,7)	1	0.528989	0.764812	3.537857
(6,8)	1	0.140685	0.328179	2.269074
average	1	0.2788	0.4656	2.9345
(1,2)	2	0.142101	0.223988	2.863277
(3,4)	2	0.359835	0.309076	2.830315
average	2	0.2510	0.2665	2.8468

where $P(i)$ is the probability of state i according to the simulation model and $Q(i)$ is the corresponding analytical approximation. The fourth and fifth column provide reference values that put column three into perspective. Column four contains the Kullback-Leibler divergence $D_{\mathrm{LK}}(P \parallel \text{marginals})$ between the simulation model and an approximation that is obtained by multiplying its respective one-dimensional marginals, which are estimated via simulation. The fifth column shows the Kullback-Leibler divergence $D_{\mathrm{LK}}(P \parallel \text{uniform})$ between the simulation model and a uniform approximation. The following observations can be made.

- The analytical model clearly outperforms the uniform approximation, meaning that it provides useful information beyond the completely uninformed case.
- The analytical model also outperforms the marginal-based approximation on average for all distances, meaning that the analytical model captures relevant dependency information.
- The marginal-based approximation improves as the distance gets larger. This is consistent with the traffic modeling intuition that queue dependencies decrease with spatial distance.
- The performance of the analytical model exhibits the sharpest reduction in quality when going from distance 0 to distance 1. This is plausible because for a pair of queues with distance 0 there exists a subnetwork that contains joint distributional information for both queues.

For each queue pair, the detailed statistics display overall the same trends; the only exception to this rule are queues 3 and 4 (second last row), for which the marginal approximation performs better than the proposed model. An inspection of the corresponding distribution plots in the last row of Figure 6 suggests that this is owed to an imperfect approximation of the two probability peaks at (10, 0) and (0, 10).

Overall, these experiments demonstrate that the proposed approximation model captures most dependency structure in a fairly ill-behaved test case that is characterized by a complex multi-modal joint distribution. The approximation model is computed using four subnetwork approximations, with each subnetwork consisting of 3 queues. Given a space capacity of 10 vehicles per queue, this implies an overall memory requirement of

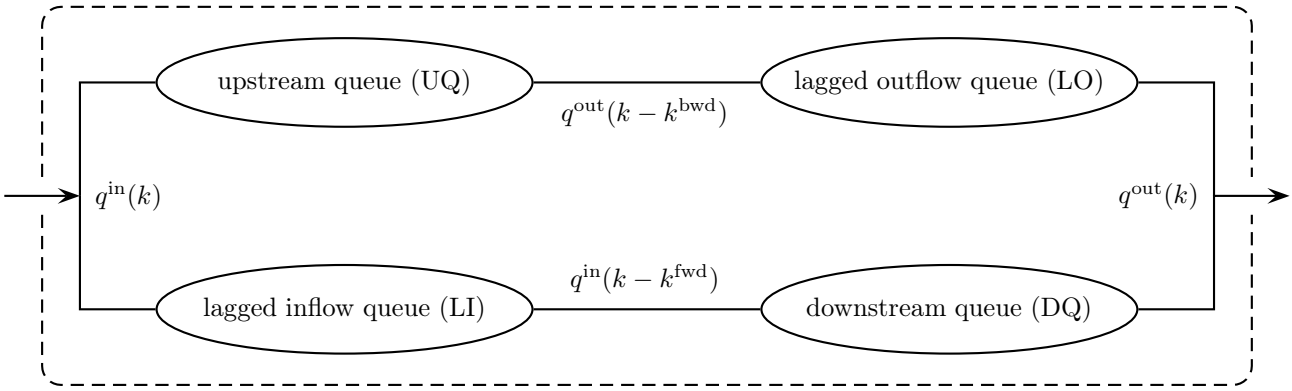


Figure 9: Link composed of four queues

$4 \times 11^3 = 5324$ numbers. Given the full state space size of $11^8 = 214'358'881$, this means a reduction down to less than 0.0025 percent. The following section puts this approximation model into concrete use for the development of a network SLTM.

3 Road network model

This section deploys the previously developed queueing network model to specify a SLTM for vehicular road network traffic. The model is developed in Section 3.1, its numerical solution is discussed in Section 3.2, and experimental illustrations are given in Section 3.3.

3.1 Model formulation

This work relies on the link model of Osorio and Flötteröd (2015), which realistically captures stochastic kinematic waves *within* a link. It is briefly reviewed in Section 3.1.1. For a detailed description of the derivation of the model, the reader is referred to Osorio and Flötteröd (2015). Sections 3.1.2 and 3.1.3 explain how the previously developed queueing network model can be used to consistently combine these link models into a linear and a general-topology network model, respectively.

3.1.1 Link model

The notation used here differs slightly from that in Osorio and Flötteröd (2015).

The link model considers an isolated link (i.e., a road segment) with a triangular density-flow fundamental diagram. Stochasticity is modeled in the arrival process to the upstream end of the link and the departure process from its downstream end. The model is parametrized by the link's free flow velocity, backward wave speed, flow capacity, jam density and length. It is a continuous-space discrete-time model that uses four finite (space) capacity Markovian queues to describe the boundary conditions the link provides to both its upstream and its downstream interface (a node in a network context).

The *downstream queue* DQ contains the number of vehicles that are ready to leave the link, constituting the boundary condition the link provides to its downstream node. The *lagged inflow* queue LI contains the total number of vehicles that have entered the link but, due to the finite link traversal speed, do not yet affect its downstream boundary condition. The sum of LI and DQ can hence be interpreted as the sum of "vehicles moving on the link" and "vehicles queueing at the downstream end of the link", yielding the total number of vehicles on the link.

The number of vehicles contained in the *upstream queue* UQ is such that the remaining space available in this queue represents the space available for vehicles entering the link, constituting the boundary condition it provides to its upstream node. The *lagged outflow* queue LO keeps track of how many vehicles have left the link but, due to the finite backward wave speed, do not yet affect its upstream boundary condition. This means that LO does not contain vehicles but what could be called "vehicle departure events" or "spaces about to become available upstream". The interplay of UQ and LO is such that UQ may contain more vehicles than what the link physically contains (because the effect of vehicles having recently left the link is not yet observable at its upstream and), in which case LO keeps track of this surplus.

Figure 9 illustrates the configuration of these queues within a link. Using k as the discrete time index, k^{fwd} (resp. k^{bwd}) is the number of time steps it takes a forward (resp. backward) kinematic wave to traverse the link. The link's in- and outflow rates are denoted by q^{in} and q^{out}, respectively.

Table 5: Transition rates between queues DQ, LO, UQ, LI. Only changed final states are indicated.

initial state m	**final state n**	**rate $t^{\mathbf{n}}_{\mathbf{m}}(k)$**	**condition**
dq, lo, uq, li	$uq+1, li+1$	$\gamma(k)$	$uq < \ell$
– " –	$li-1, dq+1$	$\mu^{\mathrm{LI}}(li;k)$	$li > 0$
– " –	$dq-1, lo+1$	$\delta(k)$	$dq > 0$
– " –	$lo-1, uq-1$	$\mu^{\mathrm{LO}}(lo;k)$	$lo > 0$

The total number of vehicles in the link can be either expressed as the sum of vehicles in DQ plus those in LI (having entered the link but not yet entered DQ) or as those in UQ minus those in LO (having left the link but not yet been taken out of UQ). Denoting by the italic symbols DQ (LO, UQ, LI) the stochastic number of vehicles in DQ (LO, UQ, LI), one hence has

$$DQ + LI = UQ - LO. \tag{21}$$

This linear dependence implies that the state of the link can be expressed by any three out of these four queues. Since the selection of which queue to leave out is arbitrary and would create the notational overhead of expressing one queue state through the remaining three, the state of the link model is in the following expressed through all four queues, keeping the linear dependence (21) in mind.

Let k be the current time step index, h the duration of a time step, and ℓ the space capacity of the link (and of each single queue it contains). Denoting by dq, lo, uq, li concrete realizations of DQ, LO, UQ, LI that comply with (21), Table 5 enumerates the rates at which transitions between these queue states occur, with "– " –" meaning "the same entry as in the row immediately above". The first (resp. second) column of Table 5 represent the initial (resp. new) state, with unchanged queue states being not repeated in the second column. The third column represents the correspoding transition rate; note that this rate is time-dependent, as described below. The fourth column represents the condition on the initial state under which this transition can take place.

- The first row of the table describes arrivals to the link. They occur with rate $\gamma(k)$ and may enter the link as long as $uq < \ell$, i.e. they may enter as long as the number of vehicles in UQ is below the space capacity ℓ.

- The second row describes flow transmissions from LI to DQ. They are transmitted with rate

$$\mu^{\mathrm{LI}}(li;k) = \frac{li}{h} \cdot \frac{q^{\mathrm{in}}(k-k^{\mathrm{fwd}})}{\sum_{j=1}^{k^{\mathrm{fwd}}} q^{\mathrm{in}}(k-j)}, \tag{22}$$

  and this can occur as long as LI is nonempty ($li > 0$). This expression combines two ingredients. First, it evaluates *lagged* link inflows. This captures the finite propagation speed of kinematic forward waves. Second, it conditions on the concrete realization li of the number of vehicles in the LI queue. In combination, this allows to keep track of the concrete distribution of flow having entered LI in past time steps. Observing that the expected state of LI represents the accumulation of the link inflows during the last k^{fwd} time steps, i.e. $\mathrm{E}\{LI(k)\} = h\sum_{j=1}^{k^{\mathrm{fwd}}} q^{\mathrm{in}}(k-j)$, it follows from (22) that $\mathrm{E}\{\mu^{\mathrm{LI}}(LI;k)\} = q^{\mathrm{in}}(k-k^{\mathrm{fwd}})$.

- Row three describes departures from the link, which occur at rate $\delta(k)$ as long as DQ is nonempty.

- The last row describes how lagged link exits affect UQ, i.e. how a space becomes available at the upstream end of the link. This is not modeled as a flow *transition* but by a joint reduction of LO and UQ. It occurs at rate

$$\mu^{\mathrm{LO}}(lo;k) = \frac{lo}{h} \cdot \frac{q^{\mathrm{out}}(k-k^{\mathrm{bwd}})}{\sum_{j=1}^{k^{\mathrm{bwd}}} q^{\mathrm{out}}(k-j)}. \tag{23}$$

  The interpretation of this equation is symmetric to that of (22), only that one now aims at capturing kinematic backward waves. One has $\mathrm{E}\{\mu^{\mathrm{LO}}(LO;k)\} = q^{\mathrm{out}}(k-k^{\mathrm{bwd}})$.

Some intuition for how this specification relates to the LTM of Yperman et al. (2006) is subsequently developed. Link boundary conditions are updated in Yperman et al. (2006) according to formulae that involve (i) differences

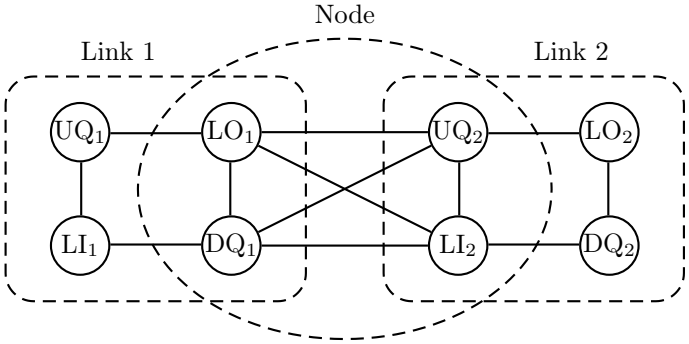


Figure 10: Queue composition of two-link tandem network

of instantaneous cumulative link inflows and time-lagged cumulative link outflows to capture the upstream boundary conditions of a link, and (ii) differences of time-lagged cumulative link inflows and instantaneous cumulative link outflows to capture the downstream boundary conditions of a link. The number of vehicles in UQ and in DQ of the present model represent stochastic versions of each of these differences, which are computed by feeding stochastic link in- and outflows with suitable time lags into these queues. Osorio and Flötteröd (2015) further demonstrate that introducing the supplementary LI and LO queues allows to derive the *joint* distribution of the link's up- and downstream conditions.

This completes the summary of the single-link model of Osorio and Flötteröd (2015).

3.1.2 Linear network model

To connect multiple link models into a network topology, one needs to add a node (i.e., intersection) model that describes how vehicles arriving at the end of a link and intending to continue their travel into a certain downstream link move across the corresponding link boundaries.

This means that vehicles can enter a link now in two different ways: from outside of the network or from an upstream link within the network. Similarly, they can leave a link either to outside of the network or to a downstream link within the network. Where exactly along a link network entries and exits occur is a matter of specification; the convention adopted here is that vehicles are inserted into the network at the upstream end of a link and that they are taken out of the network at the downstream end of a link. Geometrically, the entrance and exit locations hence coincide with the link's up- and downstream node, even though entrances and exits are defined in a link-specific manner.

Given that the previously reviewed link model is rooted in a queueing theoretical specification, the interactions between adjacent links across a node are also specified at the level of the involved queues. The model structure is illustrated in Figure 10 for a tandem configuration of two links (i.e. unidirectional roads) that are connected by a node (i.e. an intersection). Each link (upstream: link 1, downstream: link 2, both drawn as dashed rectangles) contains the four queues DQ, LO, UQ, LI that define its internal stochastic flow dynamics as reviewed in Section 3.1.1. These queues are drawn as solid circles.

The solid lines connecting the queues illustrate how they interact: Two queues are connected by a solid line if there exists a network state transition that depends on both queues or affects both queues. Table 6 displays all transitions needed in this network in order to create SLTM dynamics. The first and last block of four rows describes, now in a network context, the four types of within-link events already discussed in the explanation of Table 5. The single row in the middle block defines how the two links are joined into a network: Vehicles move at a node- and link-specific rate μ_{12} across the node given that there are vehicles available upstream to be moved ($dq_1 > 0$) and that there is space available downstream to receive these vehicles ($uq_2 < \ell_2$). This movement affects the upstream link in that it has the same effect as a single-link departure, and it affects the downstream link in that it has the same effect as a single-link arrival, with the essential difference that these two events now occur jointly.

This node model adds stochasticity and finite vehicle size in a natural way to the well-known KWM interface logic (Daganzo; 1994; Lebacque; 1996), which is deterministic and applies to a continuum vehicle flow. To clarify this, a result for the link model of Osorio and Flötteröd (2015) is recalled. They derive the following expressions for the sending flow rate S at which flow can leave a link and the receiving flow rate R at which flow can enter a link during a given short time interval of duration h, with μ being the link's flow capacity:

$$S = \min\{DQ/h,\, \mu\} \tag{24}$$

$$R = \min\{(\ell - UQ)/h,\, \mu\}\,. \tag{25}$$

Table 6: Exact transition rates in tandem network. Only changed new states are shown. The time index is omitted for better readability. Subscripts refer to the link containing the respective queue.

initial state x	**new state y**	**rate $t_{\mathbf{x}}^{\mathbf{y}}$**	**condition**
$dq_1, lo_1, uq_1, li_1; dq_2, lo_2, uq_2, li_2$	uq_1+1, li_1+1	γ_1	$uq_1 < \ell_1$
– " –	li_1-1, dq_1+1	$\mu^{\mathrm{LI}}(li_1)$	$li_1 > 0$
– " –	dq_1-1, lo_1+1	δ_1	$dq_1 > 0$
– " –	lo_1-1, uq_1-1	$\mu^{\mathrm{LO}}(lo_1)$	$lo_1 > 0$
– " –	$dq_1-1, lo_1+1, uq_2+1, li_2+1$	μ_{12}	$dq_1 > 0, uq_2 < \ell_2$
– " –	uq_2+1, li_2+1	γ_2	$uq_2 < \ell_2$
– " –	li_2-1, dq_2+1	$\mu^{\mathrm{LI}}(li_2)$	$li_2 > 0$
– " –	dq_2-1, lo_2+1	δ_2	$dq_2 > 0$
– " –	lo_2-1, uq_2-1	$\mu^{\mathrm{LO}}(lo_2)$	$lo_2 > 0$

In the limiting case of $h \to 0$, this becomes

$$S = \mu \cdot \mathbf{1}(DQ > 0) \tag{26}$$
$$R = \mu \cdot \mathbf{1}(UQ < \ell) \tag{27}$$

with $\mathbf{1}(\cdot)$ being the indicator function. Concatenating now an upstream link 1 and a downstream link 2 with respective flow capacity μ_1 and μ_2 and deploying the usual KWM interface logic, the stochastic flow is given by

$$Q_{12} = \min\{S_1, R_2\}. \tag{28}$$

Substituting (26) and (27) and noting that the resulting expression is zero unless both involved indicators are one yields

$$Q_{12} = \min\{\mu_1, \mu_2\} \cdot \mathbf{1}(DQ_1 > 0) \cdot \mathbf{1}(UQ_2 < \ell_2). \tag{29}$$
$$\Rightarrow \quad \mathrm{E}\{Q_{12}\} = \min\{\mu_1, \mu_2\} \cdot \Pr(DQ_1 > 0, UQ_2 < \ell_2) \tag{30}$$

where the subscripts 1 and 2 refer to the respective links and $\mu_{12} = \min\{\mu_1, \mu_2\}$ can now be identified as the interface flow capacity. The expected interface flow (30) coincides with the expected node transition rate in Table 6.

Given Table 6, the stochastic traffic flow dynamics on this network can be evaluated using (2). In order to tackle the exponential complexity of this equation, a suitable subnetwork decomposition is needed. This subnetwork decomposition is indicated in Figure 10 by the three regions circumscribed by dashed lines: two *link subnetworks* and one *node subnetwork*. Inspecting Figure 10 reveals that this subnetwork decomposition is *triangle-free* (Definition 2). Further, all queues refered to in every single line of Table 6 can be inscribed in a single subnetwork (the first block of rows into the subnetwork of link 1, the second block into the node subnetwork, and the last block into the subnetwork of link 2), leading to the conclusion that this specification allows for *instantaneous local transitions only* (Definition 3). All necessary prerequisites to deploy the subnetwork decomposition model (13), (14) are hence satisfied.

The local transition rates (Definition 4) necessary to evaluate (13) are given in Table 7. The first and second column contain the initial state of the considered subnetwork and its neighborhood. The third and fourth column show the corresponding states arising after the transition. Empty fields mean that the corresponding subnetwork state is not changed by the respective transition. Column five displays the rate at which this transition occurs, given that the condition in column six is fulfilled. The rows are as follows.

- The first block of rows describes all events affecting the subnetwork of link 1. This means that the rates $t_{\mathbf{m},\mathbf{r}}^{\mathbf{n},\mathbf{s}}$ given here correspond to $t_{\mathbf{m},\mathbf{r}}^{\mathbf{n},\mathbf{s}}(\mathcal{V}(\text{subnetwork of link 1}))$ in (13). This subnetwork overlaps with that of the node; its neighborhood queues are hence UQ_2 and LI_2 (i.e. the queues of the node subnetwork that are not already contained in the link 1 subnetwork). The rows in this block describe, from top to bottom:

initial state $(\mathbf{m},\mathbf{r})$		**final state** $(\mathbf{n},\mathbf{s})$		**rate** $t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}$	**condition**
link 1 $(\mathbf{m})$	**node** $(\mathbf{r})$	**link 1** $(\mathbf{n})$	**node** $(\mathbf{s})$		
dq_1, lo_1, uq_1, li_1	uq_2, li_2	uq_1+1, li_1+1		γ_1	$uq_1 < \ell_1$
– " –	– " –	li_1-1, dq_1+1		$\mu^{\mathrm{LI}}(li_1)$	$li_1 > 0$
– " –	– " –	dq_1-1, lo_1+1		δ_1	$dq_1 > 0$
– " –	– " –	lo_1-1, uq_1-1		$\mu^{\mathrm{LO}}(lo_1)$	$lo_1 > 0$
– " –	– " –	dq_1-1, lo_1+1	uq_2+1, li_2+1	μ_{12}	$dq_1 > 0$, $uq_2 < \ell_2$
node $(\mathbf{m})$	**link 1, link 2** $(\mathbf{r})$	**node** $(\mathbf{n})$	**link 1, link 2** $(\mathbf{s})$		
dq_1, lo_1, uq_2, li_2	uq_1, li_1, dq_2, lo_2	dq_1+1	li_1-1	$\mu^{\mathrm{LI}}(li_1)$	$li_1 > 0$
– " –	– " –	dq_1-1, lo_1+1		δ_1	$dq_1 > 0$
– " –	– " –	lo_1-1	uq_1-1	$\mu^{\mathrm{LO}}(lo_1)$	$lo_1 > 0$
– " –	– " –	$dq_1-1, lo_1+1, uq_2+1, li_2+1$		μ_{12}	$dq_1 > 0$, $uq_2 < \ell_2$
– " –	– " –	uq_2+1, li_2+1		γ_2	$uq_2 < \ell_2$
– " –	– " –	li_2-1	dq_2+1	$\mu^{\mathrm{LI}}(li_2)$	$li_2 > 0$
– " –	– " –	uq_2-1	lo_2-1	$\mu^{\mathrm{LO}}(lo_2)$	$lo_2 > 0$
link 2 $(\mathbf{m})$	**node** $(\mathbf{r})$	**link 2** $(\mathbf{n})$	**node** $(\mathbf{s})$		
dq_2, lo_2, uq_2, li_2	dq_1, lo_1	uq_2+1, li_2+1		γ_2	$uq_2 < \ell_2$
– " –	– " –	li_2-1, dq_2+1		$\mu^{\mathrm{LI}}(li_2)$	$li_2 > 0$
– " –	– " –	dq_2-1, lo_2+1		δ_2	$dq_2 > 0$
– " –	– " –	lo_2-1, uq_2-1		$\mu^{\mathrm{LO}}(lo_2)$	$lo_2 > 0$
– " –	– " –	uq_2+1, li_2+1	dq_1-1, lo_1+1	μ_{12}	$dq_1 > 0$, $uq_2 < \ell_2$

Table 7: Local transition rates for subnetworks in 2-link tandem network

  - arrival from outside of the network to link 1;
  - advancement of a vehicle in LI_1 into DQ_1;
  - a vehicle leaving link 1 out of the network;
  - a "vehicle departure event" leaving LO_1 and releasing a space in UQ_1;
  - a vehicle leaving link 1 and continuing into link 2.

- The second block of rows describes all events affecting the node subnetwork. This means that the rates $t^{\mathbf{n,s}}_{\mathbf{m,r}}$ given here correspond to $t^{\mathbf{n,s}}_{\mathbf{m,r}}(\mathcal{V}(\text{node subnetwork}))$ in (13). This subnetwork overlaps with those of both links; its neighborhood queues are hence UQ_1 and LI_1 (the queues of the link 1 subnetwork that are not already contained in the node network), and DQ_2 and LO_2 (the queues of the link 2 subnetwork that are not already contained in the node network). The rows in this block describe, from top to bottom:
  - advancement of a vehicle in LI_1 into DQ_1;
  - departure out of the network from the downstream end of link 1;
  - a "vehicle departure event" leaving LO_1 and releasing a space in UQ_1;
  - a vehicle leaving link 1 and continuing into link 2;
  - arrival from outside of the network to link 2;
  - advancement of a vehicle in LI_2 into DQ_2;
  - a "vehicle departure event" leaving LO_2 and releasing a space in UQ_2.

- The third block of rows describes all events affecting the subnetwork of link 2. This means that the rates $t^{\mathbf{n,s}}_{\mathbf{m,r}}$ given here correspond to $t^{\mathbf{n,s}}_{\mathbf{m,r}}(\mathcal{V}(\text{subnetwork of link 2}))$ in (13). This subnetwork overlaps with that of the node; its neighborhood queues are hence DQ_1 and LO_1. The rows in this block describe the same type of events for link 2 as the rows in the first block for link 1.

One observes that events affecting more than one subnetwork are repeated in the definition of the local transition rates of each involved subnetwork. This consequence of Definition 4 reflects the fact that subnetworks may overlap and is essential for capturing stochastic dependency between subnetworks.

This completes the specification of the proposed network SLTM for a two-link tandem network. To use this framework for the modeling of general road network topologies, the following is necessary.

1. Every road direction is represented by a four-queue link model. One link subnetwork is defined for every link.
2. One node subnetwork is defined for every road intersection. It comprises DQ and LO of all upstream (ingoing) links and UQ, LI of all downstream (outgoing) links of that node.
3. Concrete transition rates are defined for each node subnetwork. These rates model the concrete intersection under consideration.

Items 1 and 2 imply that every sequence of overlapping subnetworks alternates between *link subnetworks* and *node subnetworks*. This means that all subnetworks adjacent to a *node subnetwork* are *link subnetworks*, and vice versa. As a consequence, the resulting subnetwork structure is *triangle-free*. Item 3 requires to specify a stochastic node model that, for a general network, may allow for an arbitrary number of in- and outgoing links. The SLTM framework is flexible with respect to the concrete node model specification. An example diverge and merge node model are subsequently developed.

3.1.3 General network model

Every node specification must allow for *instantaneous local transitions only* (Definition 3). This requirement is automatically satisfied if the flows across a node depend only on the corresponding boundary conditions of the adjacent links, as in standard KWM theory.

In an node with more than one up- or downstream link, every vehicle moving across that node comes from one particular upstream link or moves towards one particular downstream link. Given finite vehicle sizes, crossing the node takes finite time, and the information of where a vehicle comes from or where it goes does not change while the vehicle advances. Capturing this information in the SLTM would require to introduce corresponding state variables because the model is Markovian along the time-line. The subsequently presented merge and diverge model aim at simplicity and approximate node flows without such a state space expansion.

Let I and J be the number of the node's in- and outgoing links. As a general convention, ingoing (upstream) links are indexed by the symbol i, outgoing (downstream) links by symbol j, and the symbol l is used when up- or downstream information does not play a role or when a secondary index is necessary.

Table 8: Transition rates from upstream link i to downstream link j across different node types

node type	**transition rate**	**condition**
straight	$\min\{\mu_i, \mu_j\}$	$dq_i > 0$ and $uq_j < \ell_j$
diverge	$p_{ij} \min\left\{\mu_i, \min\limits_{\{l \text{ downstr.}\}} \left\{\frac{\mu_l}{p_{il}}\right\}\right\}$	$dq_i > 0$ and $\forall l$ downstr.:($uq_l < \ell_l$ or $p_{il} = 0$)
merge	$\alpha_i \left(\sum\limits_{\{l \text{ upstr. with } dq_l>0\}} \frac{\alpha_l}{\min\{\mu_l, \mu_j\}}\right)^{-1}$	$dq_i > 0$ and $uq_j < \ell_j$

General diverge

A general diverge node has $I = 1$ upstream links and $J > 1$ downstream links. The turning probability from the unique upstream link i into downstream link j is denoted by p_{ij}. Conservation of turning fractions (meaning here that the ratios of transition rates are equal to the corresponding turning probability ratios, cf. Tampere et al. (2011)) is ensured by declaring the diverge as blocked (i.e. unable to transmit any flow) whenever the UQ of a downstream link j with $p_{ij} > 0$ is full. (Relaxing this condition, i.e. sending flow into a non-full downstream link while another downstream link is full would require the aforementioned state space extension to keep track of of the destination link of vehicles queueing upstream.)

Concrete transition rates are adopted from the broadly used diverge model of Daganzo (1995a), in that the node flow is maximized subject to the following constraints: The outflow out of the upstream link i does not exceed its flow capacity μ_i; the inflow into every downstream link j does not exceeds its flow capacity μ_j; turning fractions are preserved. Given that the diverge is not blocked, the flow rate from upstream then becomes $\min\left\{\mu_i, \min_{\{l \text{ downstream}\}} \left\{\frac{\mu_l}{p_{il}}\right\}\right\}$, which is distributed according to the turning probabilities p_{ij} into the respective downstream links. This model follows from the same derivation as given in Daganzo (1995a), only that the SLTM's discrete vehicle representation implies that the rate at which an upstream link can send (resp. a downstream link can receive) is either zero (if there is no vehicle resp. space available) or the link's flow capacity μ (if there is at least one vehicle resp. space available).

General merge

A general merge node has $I > 1$ upstream links and $J = 1$ downstream links. The flow capacity between upstream link i and the unique downstream link j is $\min\{\mu_i, \mu_j\}$, meaning that the expected transition time of a single vehicle from i to j is $1/\min\{\mu_i, \mu_j\}$. Every upstream link i receives a strictly positive priority parameter α_i that guides the way in which possible competition for downstream capacity is resolved. Letting the set $C = \{i \text{ upstream}: dq_i > 0\}$ contain all upstream links that currently compete for downstream capacity, the probability that link $i \in C$ wins this competition is set to $\alpha_i / \sum_{j \in C} \alpha_j$.

The probability that a vehicle currently moving across the node comes from upstream link $i \in C$ is approximated by the probability $\alpha_i / \sum_{l \in C} \alpha_l$ that a vehicle from this link would win an instantaneous competition. The expected time it takes the currently advancing vehicle, regardless of where it comes from, to move across the node is hence approximated by $\sum_{i \in C} \frac{\alpha_i}{\sum_{l \in C} \alpha_l} \cdot \frac{1}{\min\{\mu_i, \mu_j\}}$. Inverting this expression yields the total flow rate $\sum_{l \in C} \frac{\alpha_l}{\sum_{i \in C} \alpha_i / \min\{\mu_i, \mu_j\}} = \sum_{i \in C} \frac{\alpha_i}{\sum_{l \in C} \alpha_l / \min\{\mu_l, \mu_j\}}$. The last expression results from exchanging the l and i summation indices; the purpose of this is merely to subsequently follow the convention that i refers to an upstream link. Given that there is space available downstream, i.e. $uq_j > 0$, the resulting flow transmission rate between upstream link $i \in C$ and downstream link j is then set to the corresponding addend in the last sum, i.e. to $\frac{\alpha_i}{\sum_{l \in C} \alpha_l / \min\{\mu_l, \mu_j\}}$.

Table 8 summarizes the transition rates across the different types of nodes discussed in this article. The transition rates necessary to specify a full network SLTM that contains these nodes in arbitrary topology is given in Table 9. The presentation avoids redundancies and is hence somewhat more compact than in the earlier tables. It consists of two blocks of rows, the first one defining the transition rates for a link subnetwork and the second one defining the transition rates for a node subnetwork. The first column of Table 9 indicates the type of considered transition. The notation of the following columns is such that they can be immediately inserted into the general subnetwork dynamics (13), (14), which require defining the transition rates $t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{V}(S))$ for each subnetwork S with $\mathbf{m}, \mathbf{n} \in \mathfrak{N}(\mathcal{V}(S))$ being states of subnetwork S and $\mathbf{r}, \mathbf{s} \in \mathfrak{N}(\partial\mathcal{V}(S)))$ being states

of its neighborhood $\partial\mathcal{V}(S)$. Specifically, the second column indicates those components of the initial state $\mathbf{m}, \mathbf{r}$ that change during the transition. The third column indicates those components of the final state $\mathbf{n}, \mathbf{s}$ that have changed during the transition. Column four shows the rate $t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{V}(S))$ at which the transition takes place, given that the condition in column five is satisfied. For brevity, some entires in column four and five refer back to Table 8. The following specifications are given for a link subnetwork.

- The first four rows of the first block refer to events that are fully contained in the link subnetwork: departures out of the network, arrivals from outside of the network, transitions from LI to DQ, transitions from LO to UQ.
- Row five (resp. six) of the first block indicates what happens when a vehicle enters (resp. leaves) the considered link from an upstream link i (resp. to a downstream link j). Here, states in the neighborhood of the considered link subnetwork are also changed; these states refer to downstream boundary conditions of the sending upstream link i (resp. to upstream boundary conditions of the receiving downstream link j).

The following specifications are given for a node subnetwork.

- The first three rows of the second block refer to events that are fully contained in the node subnetwork: departure out of the network from an upstream link (which only affects the downstream boundary conditions of that link, which are part of the node subnetwork), arrival to the network in a downstream link (which only affects the upstream boundary conditions of that link, which are part of the node subnetwork), and a transition from an up- to a downstream link (which also only affects those parts of the involved links that are part of the node subnetwork).
- Rows four and five of the second block refer to transitions from LI to DQ and from LO to UQ in an upstream link of the node. Since UQ and LI of that link are not part of the node subnetwork, the corresponding subnetwork neighborhood states are also changed.
- The last two rows of the second block describe the same transition types as in the previous item, but now in a downstream link of the node. Symmetrically to the previous case, since that link's DQ and LO are not contained in the node subnetwork, the corresponding subnetwork neighborhood states are also changed.

This completes the specification of all network SLTM elements. The full network model and its numerical solution are presented in the following section.

3.2 Continuous-time network model and numerical solution

The LTM of Yperman et al. (2006) is, as well as its stochastic counterpart (Osorio and Flötteröd; 2015), specified in discrete time. The queueing subnetwork dynamics (13), (14) on which the network SLTM of the present article builds are, however, specified in continuous time.

Consistency between these two time representations is subsequently established by reformulating the stochastic LTM of Osorio and Flötteröd (2015) in continous time. For this, it is recalled that the network SLTM requires to insert the transition rates of Tables 8 and 9 into the continous-time subnetwork dynamics (13), (14). An overall continuous-time formulation hence results if all involved transition rates are defined in continuous time. The only dependencies on a discrete time formulation that can be identified refer to $\mu^{\mathrm{LI}}(li)$ and $\mu^{\mathrm{LO}}(lo)$ in Table 9, which hence are reformulated in continous time.

For this, the lagged inflow rate (22) is written as

$$\mu^{\mathrm{LI}}(li; kh) \quad = \quad li \cdot \frac{q^{\mathrm{in}}(kh - k^{\mathrm{fwd}}h)}{\sum_{j=1}^{k^{\mathrm{fwd}}} h \cdot q^{\mathrm{in}}(kh - jh)}, \tag{31}$$

with the main difference to (22) being that the discrete time index k is here replaced by discrete points kh with distance h in continuous time. The denominator of this expression is a Riemann Sum over a time-continuous inflow profile $q^{\mathrm{in}}(\tau)$ in the time interval $[\tau - \tau^{\mathrm{fwd}}, \tau]$ with $\tau = kh$ and $\tau^{\mathrm{fwd}} = k^{\mathrm{fwd}}h$. One obtains

$$\lim_{h \to 0} \mu^{\mathrm{LI}}(li; \tau) \quad = \quad li \cdot \frac{q^{\mathrm{in}}(\tau - \tau^{\mathrm{fwd}})}{\int_{\varrho=0}^{\tau^{\mathrm{fwd}}} q^{\mathrm{in}}(\tau - \varrho) d\varrho} \tag{32}$$

	event type	initial components of $\mathbf{m}$; $\mathbf{r}$	final components of $\mathbf{n}$; $\mathbf{s}$	rate $t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{V}(S))$	condition
link l subnetwork S	departure	dq_l, lo_l; $-$	$dq_l - 1, lo_l + 1$; $-$	δ_l	$dq_l > 0$
	arrival	uq_l, li_l; $-$	$uq_l + 1, li_l + 1$; $-$	γ_l	$uq_l < \ell_l$
	lagged inflow	dq_l, li_l; $-$	$dq_l + 1, li_l - 1$; $-$	$\mu_l^{\mathrm{LI}}(li_l) \rightarrow (22)$	$li_l > 0$
	lagged outflow	lo_l, uq_l; $-$	$lo_l - 1, uq_l - 1$; $-$	$\mu_l^{\mathrm{LO}}(lo_l) \rightarrow (23)$	$lo_l > 0$
	transition from upstream link i	uq_l, li_l; dq_i, lo_i	$uq_l + 1, li_l + 1$; $dq_i - 1, lo_i + 1$	$\rightarrow$ Table 8	$\rightarrow$ Table 8
	transition to downstream link j	dq_l, lo_l; uq_j, li_j	$dq_l - 1, lo_l + 1$; $uq_j + 1, li_j + 1$	$\rightarrow$ Table 8	$\rightarrow$ Table 8
node subnetwork S	departure from upstr. link i	dq_i, lo_i; $-$	$dq_i - 1, lo_i + 1$; $-$	δ_i	$dq_i > 0$
	arrival to downstr link j	uq_j, li_j; $-$	$uq_j + 1, li_j + 1$; $-$	γ_j	$uq_j < \ell_j$
	transition from link i to link j	dq_i, lo_i, uq_j, li_j; $-$	$dq_i - 1, lo_i + 1, uq_j + 1, li_j + 1$; $-$	$\rightarrow$ Table 8	$\rightarrow$ Table 8
	lagged inflow in upstr. link i	dq_i; li_i	$dq_i + 1$; $li_i - 1$	$\mu_i^{\mathrm{LI}}(li_i) \rightarrow (22)$	$li_i > 0$
	lagged outflow in upstr. link i	lo_i; uq_i	$lo_i - 1$; $uq_i - 1$	$\mu_i^{\mathrm{LO}}(lo_i) \rightarrow (23)$	$lo_i > 0$
	lagged inflow in downstr. link j	li_j; dq_j	$li_j - 1$; $dq_j + 1$	$\mu_j^{\mathrm{LI}}(li_j) \rightarrow (22)$	$li_j > 0$
	lagged outflow in downstr. link j	uq_j; lo_j	$uq_j - 1$; $lo_j - 1$	$\mu_j^{\mathrm{LO}}(lo_j) \rightarrow (23)$	$lo_j > 0$

Table 9: Transition table for general network topologies

and, by symmetrical operations,

$$\lim_{h\to 0} \mu^{\mathrm{LO}}(li;\tau) \quad = \quad lo \cdot \frac{q^{\mathrm{out}}(\tau - \tau^{\mathrm{bwd}})}{\int_{\varrho=0}^{\tau^{\mathrm{bwd}}} q^{\mathrm{out}}(\tau - \varrho) d\varrho}. \tag{33}$$

The result is an overall time-continuous model, which consists of the system of differential equations (13), (14), using the transition rates from Tables 8 and 9 in conjunction with (32), (33) instead of their discrete-time counterparts (22), (23).

The present article uses a basic Euler scheme to solve this model. Note that using the Euler scheme again means approximating (32), (33) by (22), (23). Algorithm 2 summarizes the model building and solution process.

3.3 Model validation

3.3.1 Linear network model

An experiment presented in Sumalee et al. (2011) is adopted using the proposed model. The considered network consists of two unidirectional roads in tandem, the upstream road (link 1) having four lanes and being 300 meters long and the downstream road (link 2) having three lanes and being 100 meters long. Both links have triangular density-flow fundamental diagrams with maximum speeds of 60 km/h and backward wave speeds of -20 km/h. The upstream link has a jam density (summing over all four lanes) of 600 veh/km, a resulting space capacity of 180 vehicles and a resulting flow capacity of 9000 veh/h; the downstream link has a jam density (summing over all three lanes) of 400 veh/km, a resulting space capacity of 40 vehicles and a resulting flow capacity of 6000 veh/h. The interface between the two links hence constitutes a bottleneck.

Vehicles arrive to the upstream end of link 1 at a rate of

$$\gamma_1(k) \quad = \quad \begin{cases} 3000\,\mathrm{veh/h} & \text{if } hk < 250\,\mathrm{s} \\ 8000\,\mathrm{veh/h} & \text{if } hk \geq 250\,\mathrm{s} \end{cases} \tag{34}$$

with k being the time step index and h being the time step length (0.1 seconds in the present example). Vehicles leave from the downstream end of link 2 at a departure rate of $\delta_2 = 6000\,\mathrm{veh/h}$. This tandem network can be represented by the proposed model as illustrated in Figure 10, using the subnetwork transition rates of Table 7, with the flow capacity μ_{12} of the intermediate node being set to the minimum of its up- and downstream links flow capacity, i.e. to $6000\,\mathrm{veh/h}$.

The original experiment of Sumalee et al. (2011) analyses a stochastic cell transmission model. It (i) represents the upstream link by three individual cells and (ii) models stochasticity in the supply parameters maximum speed, backward wave speed, and jam density. Differently from this, the analysis presented here (i) represents the links without any cell discretization and (ii) models stochasticity in the network arrivals, inter-link transitions, and network departures. The comparability of these case studies is therewith limited; the primary objective of the present study is to illustrate the proposed model. The results are shown in Figures 11a to 11c.

The solid lines in Figure 11a and 11b show the expected total number $\mathrm{E}\{N_1\}$ and $\mathrm{E}\{N_2\}$ of vehicles in link 1 and 2, respectively, over simulation time. (N is computed as the sum of LI and DQ, cf. 21.) The dashed lines indicate the $\pm$ one standard deviation band around these means.

The dynamics of the (distribution of the) number N_1 of vehicles on link 1 are as follows. During the first 250 seconds, the average network inflow is below the bottleneck flow capacity, leading to free-flow conditions. Once the bottleneck activates, spillback arises and the number of vehicles on link 1 increases. The variance of the number of vehicles grows with their expected number. Before the bottleneck activates, the ratio of $\mathrm{VAR}\{N_1\}/\mathrm{E}\{N_1\}$ reaches a value of around 1.1. After the bottleneck has activated and stationary overcritical conditions have been attained, a ratio of $\mathrm{VAR}\{N_1\}/\mathrm{E}\{N_1\} \approx 0.63$ is attained.

Link 2 experiences undercritial conditions until the bottleneck at its upstream end activates; $\mathrm{VAR}\{N_2\}/\mathrm{E}\{N_2\}$ reaches up to this point in time a value of about 1.2. After activation of the bottleneck, one observes an overshoot in the expectation of N_2 before the link reaches marginally critical conditions (inflow rate equals outflow capacity), still with $\mathrm{VAR}\{N_2\}/\mathrm{E}\{N_2\} \approx 1.2$. It can be ascertained that this overshoot is neither a numerical artifact nor a consequence of the way in which the subnetwork decomposition approximates network-wide dependencies; this phenomenon has been confirmed through Monte-Carlo experiments with the same system.

Algorithm 2 Network SLTM construction and simulation logic

1. Construct the network representation.

   (a) Build one link subnetwork per direction of a homogeneous road segment.

      i. The subnetwork consists of one UQ, one LI, one DQ and one LO.
      ii. Set the space capacity ℓ of all queues to the road segment's space capacity.
      iii. Set the forward lag τ^{fwd} to the road segments free-flow travel time.
      iv. Set the backward lag τ^{bwd} to the traversal time of a kinematic backward wave.
      v. Set the flow capacity μ of UQ and LI to the road segment's inflow capacity.
      vi. Tet the flow capacity μ of DQ and LO to the road segments outflow capacity.
      vii. Set the arrival rate γ and departure rate δ from/to outside of the network.

   (b) Build one node subnetwork per interface between two or more homogeneous road segments.

      i. The subnetwork consists of
         - one DQ, one LO per upstream road segment and
         - one UQ, one LI per downstream road segment,

         with all queueing parameters being taken over from the respective link subnetworks.
      ii. If diverge node, set turning probabilities $\{p_{ij}\}$.
      iii. If merge node, set inflow priorities $\{\alpha_i\}$.

2. Initialize solver and model.

   - Set a simulation time step size h.
   - Set initial subnetwork distributions Φ that are consistent across overlapping subnetworks. To start with an empty network, set the probability mass of all subnetwork distributions Φ to the state representing all-empty queues.

3. For $k = 0, 1, 2, \ldots$, iterate.

   (a) Set the current model time to $\tau = kh$.

   (b) Update time-dependent network parameters.

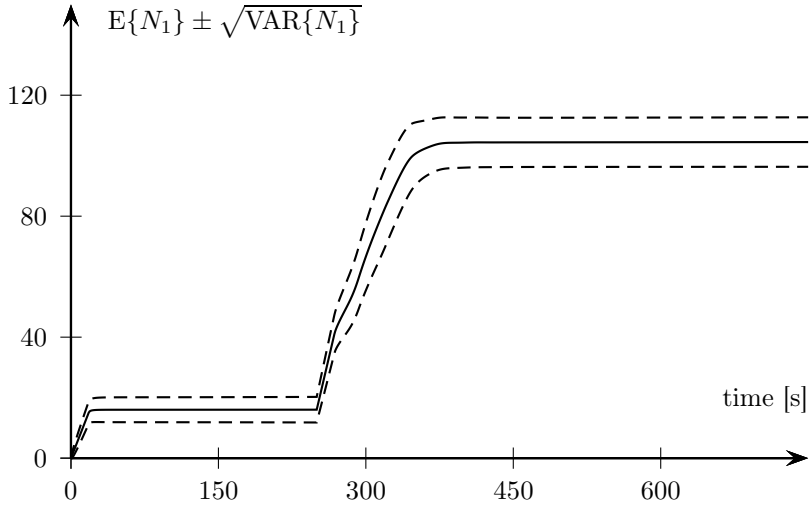
      - For link subnetworks: flow capacities $\mu(\tau)$; arrival and departure rates $\gamma(\tau)$ and $\delta(\tau)$.
      - For node subnetworks: turning probabilities $\{p_{ij}(\tau)\}$ and inflow priorities $\{\alpha_i(\tau)\}$.

   (c) Obtain node transition rates from Table 8.

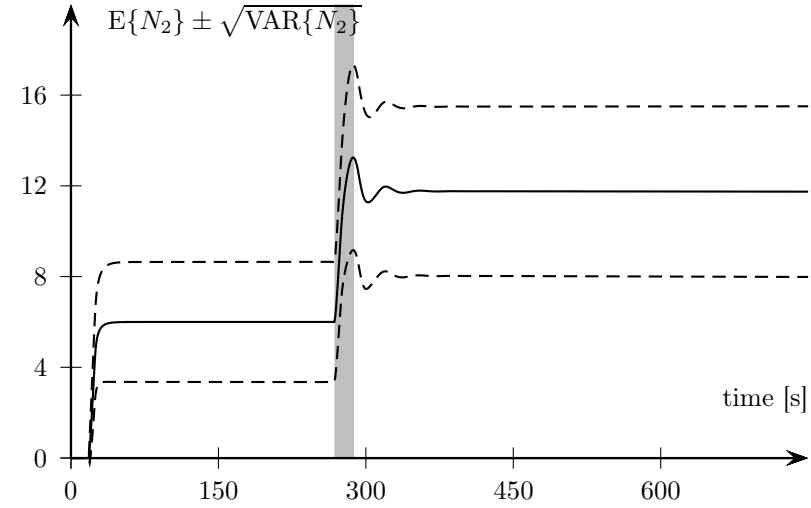
   (d) Obtain within-link transition rates from (22), (23).

   (e) Obtain subnetwork transition rates from Table 9.

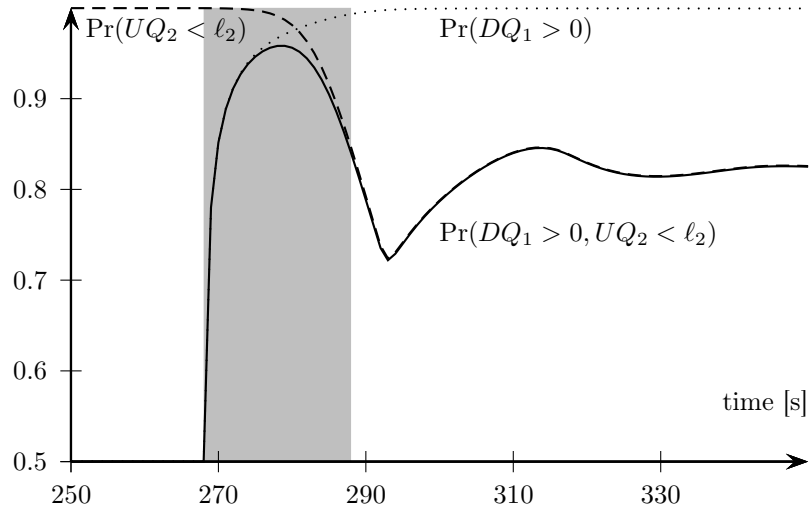
   (f) Compuate subnetwork state distributions $\Phi(\tau + h)$ by applying the Euler scheme to the system (13), (14), using current subnetwork transition rates and state distributions $\Phi(\tau)$.



(a) Statistics of number of vehicles N_1 on link 1.



(b) Statistics of number of vehicles N_2 on link 2.



(c) Queue states at the bottleneck.

Figure 11: Bottleneck experiment

To identify the mechanisms that underly this phenomenon, recall that the expected value of the stochastic flow Q_{12} through the bottleneck between link 1 and 2 is given by

$$\mathrm{E}\{Q_{12}\} \quad = \quad \mu_{12} \Pr(DQ_1 > 0, UQ_2 < \ell_2). \tag{35}$$

This means that there is in terms of *expected* bottleneck throughput no crisp difference between under- and overcritical conditions at the interface: Even in free-flow conditions the downstream conditions $\Pr(UQ_2 < \ell_2)$ take effect, and even in congested conditions the upstream conditions $\Pr(DQ_1 > 0)$ play a role. This phenomenon is not in contradiction to what one would expect based on the invariance principle (Lebacque and Khoshyaran; 2005)[1] because Figure 11c merely displays a dependence of expected flows on the *probability* of different boundary conditions in a stochastic model, whereas the invariance principle applies to the dependence of deterministic flows on deterministic boundary conditions. Indeed, as long as a flow transmission is possible at all (at least one upstream vehicle and one downstream space), the SLTM prescribes a transmission rate that is independent of how many upstream vehicles or downstream spaces are available, cf. Table 8.

Letting $Q^{\mathrm{in}}(\tau)$, $Q^{\mathrm{out}}(\tau)$ and $N(\tau)$ be the stochastic inflow, outflow, and total number of vehicles in an initially empty link, one further has

$$N(t) \quad = \quad \int_{\varrho=0}^{t} \left[Q^{\mathrm{in}}(\varrho) - Q^{\mathrm{out}}(\varrho)\right] d\varrho \tag{36}$$

$$\Rightarrow \quad \mathrm{E}\{N(t)\} \quad = \quad \int_{\varrho=0}^{t} \left[\mathrm{E}\{Q^{\mathrm{in}}(\varrho)\} - \mathrm{E}\{Q^{\mathrm{out}}(\varrho)\}\right] d\varrho, \tag{37}$$

meaning that an overshoot in the *expected* flow can also be expected to be visible in the *expected* number of vehicles on the link, as observed in Figure 11b.

The transient situation at the bottleneck after the demand increase at time 250 s is subsequently of interest; this increased inflow reaches the bottleneck at time 268 s. At this time, a large amount of vehicles has just arrived upstream of the bottleneck, while downstream there still is a lot of space (low traffic). In a *deterministic* KWM, the bottleneck would now activate and as of then allow for a constant flow rate equal to its flow capacity μ_{12}. The *stochastic* model, on the other hand, allows overcritical conditions in link 2 to arise with a certain probability, meaning that the state of link 2 affects the *expected* bottleneck flow throughout. This is illustrated in Figure 11c, which shows the probability $\Pr(DQ_1 > 0, UQ_2 < \ell_2)$ as well as its marginals $\Pr(DQ_1 > 0)$ (representing upstream congestion) and $\Pr(UQ_2 < \ell_2)$ (representing downstream space) over the time interval of interest. From second 268 to approximately second 288, $\Pr(DQ_1 > 0, UQ_2 < \ell_2)$ (and hence $\mathrm{E}\{Q_{12}\}$) overshoots compared to its subsequent stationary value. This region is underlaid with a light gray rectangle. At the beginning of this time interval, one has $\Pr(DQ_1 > 0, UQ_2 < \ell_2) \approx \Pr(DQ_1 > 0)$, representing under-critical conditions. Around second 278, $\Pr(UQ_2 < \ell_2)$ starts dominating the bottleneck flow, meaning that overcritical conditions arise. But at this time, the overshoot of $\Pr(DQ_1 > 0, UQ_2 < \ell_2)$ has already reached its maximum value. Revisiting Figure 11b, where the same time interval is underlaid in light gray, one observes that the overshoot in link 2's expected number of vehicles reaches its maximum when the overshoot in $\Pr(DQ_1 > 0, UQ_2 < \ell_2)$ has ceased (at around second 288), which is sensible given that N_2 results from a time integration of Q_{12}.

The conclusions to be drawn from this experiment are nontrivial. The network SLTM, which approximates the full state space of the tandem network under consideration, reveals damped oscillations in the expected network states. These oscillations can be traced back to the blending of under- and overcritical traffic states in the computation of expected flows. It is noteworthy that the same type of oscillations has been observed in the stochastic cell transmission model (Zhong et al.; 2013), where similar explanations (blending of uncer- and overcritical conditions) have been given. It appears sensible to draw the conclusion that analyzing time-dependent mean values as if they were realizations can lead to counter-intuitive results. The proposed network SLTM enables a much richer analysis, which is yet to be fully explored.

3.3.2 General network model

This experiment illustrates the concrete diverge and merge node models of Section 3.1.3 through the network shown in Figure 12. It consists of four uni-directinal links. The double-lined links have a length of 250 m, a space capacity of 35 veh, a maximum velocity of 60 km/h, and a flow capacity of 2100 veh/h. The single-lined link has a length of 125 m, a space capacity of 17 veh, a maximum velocity of 30 km/h, and a flow capacity of

[1] Informally, the invariance principle states that the flow through an interface must (i) in uncongested conditions not be sensitive to small changes in the downstream boundary conditions and (ii) in congested conditions not be sensitive to small changes in the upstream boundary conditions.

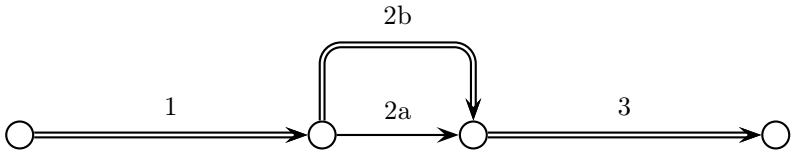


Figure 12: Test network

1680 veh/h. All links have a backwards wave speed of 20 km/h. This setting could represent an arterial bypass around a low-speed village center.

The resulting forward time lag on all links is 15 s; the backward time lag is 45 s on the high-capacity links and 22 s on the low-capacity link. The diverge turning probabilities are 50/50, this is behaviorally compatible with the observation that the free-flow travel times are identical on either routes. The merge priorities are proportional to the respective link capacities, meaning that the priority of link 2b is 1.25 times the priority of link 2a. [2]

A constant inflow of 2000 veh/h starts entering at the upstream node of the initially empty network at time zero. The system is simulated with 0.1 s time steps until it reaches near-stationary conditions after 400 seconds. (Plotting on longer time scales would merely compress the interesting transients.)

Figures 13a-d display the relative occupancy (ratio of the expected number of vehicles on a link over the respective link's space capacity) on all links; one standard deviation bands are also provided. Link 1 reacts with the previously discussed damped oscillations to the abrupt increase in arrival rate at time zero. Indeed, as also observed by (Zhong et al.; 2013), oscillating expected values appear to be triggered by rapid changes in link boundary conditions.

Focusing in this experiment on the network effects, the two parallel links 2a and 2b are considered next in Figures 13c and 13d. Their upstream diverge allocates to either link the same inflow; their downstream merge gives a higher priority to link 2b. The consequence of more yielding vehicles on link 2a is an increased probability of this link spilling back and hence reducing the throughput of its upstream diverge. This is illustrated in Figure 13e, which displays, on a logarithmic ordinate, the probability of the following events:

- $DQ_1 > 0, UQ_{2a} = \ell_{2a}, UQ_{2b} < \ell_{2b}$, meaning that a potential flow transmission from link 1 to link 2b is blocked back by link 2a.
- $DQ_1 > 0, UQ_{2a} < \ell_{2a}, UQ_{2b} = \ell_{2b}$, meaning that a potential flow transmission from link 1 to link 2a is blocked back by link 2b.

It is noteworthy that these are *joint* events involving both the up- and the downstream links of the diverge node. The possibility of spillback at the diverge means that it functions as a bottleneck, which can be read out of Figures 13a and b, where one observes congestion on the ingoing link 1 is higher than on the outgoing link 3.

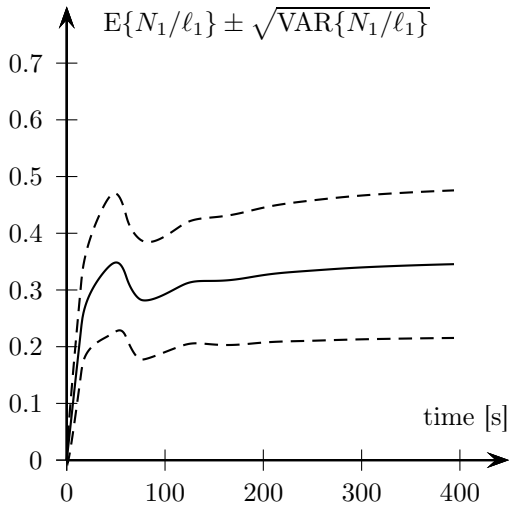
In brief summary, this experiment demonstrates that the proposed SLTM is capable of modeling sensible dynamic and stochastic flow patterns in general network topologies.

4 Summary and outlook

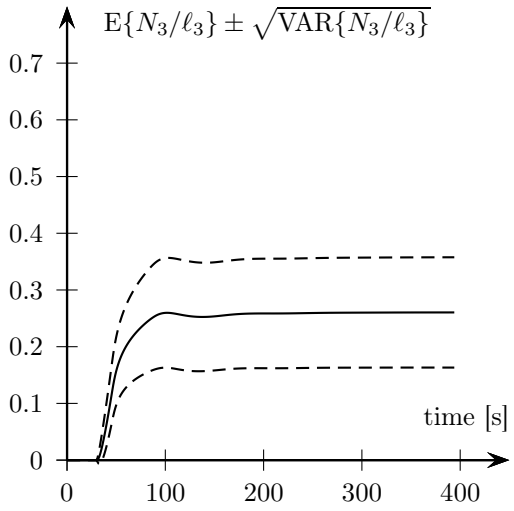
This article presents a new stochastic dynamic model of vehicular network flows. The model is rooted in finite capacity queueing theory in that all flows and road (boundary) states at the *road* network level are represented by transition rates and queue states in an underlying *queueing* network. The result is a stochastic link transmission model (SLTM) for networks.

To capture stochastic dependencies between queues, a new analytical approximation of the transient joint queue-length distributions in finite capacity Markovian networks is introduced. The approach is based on a network decomposition into overlapping subnetworks. The temporal derivative of the joint queue-length distribution of a given subnetwork is computed exclusively from (i) the joint distribution of that subnetwork and (ii) the joint distributions of all subnetworks that overlap with it. The decomposition approach is proven to be self-consistent in the sense that if any two subnetwork distributions have identical marginals for their common set of queues at some point in time, then these marginals remain identical across all other times.

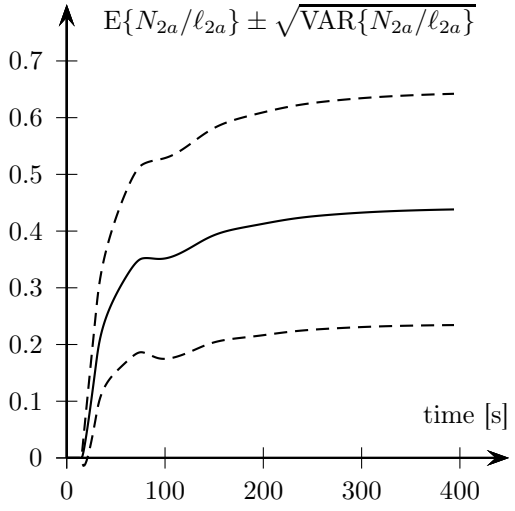
[2] The maximum of link 2a's triangular fundamental diagram allows for a maximum flow rate of about 1654 veh/h, which is slightly below that link's flow capacity parameter of 1680 veh/h. In consequence, the latter parameter only takes effect in the capacity-proportional priority setting.



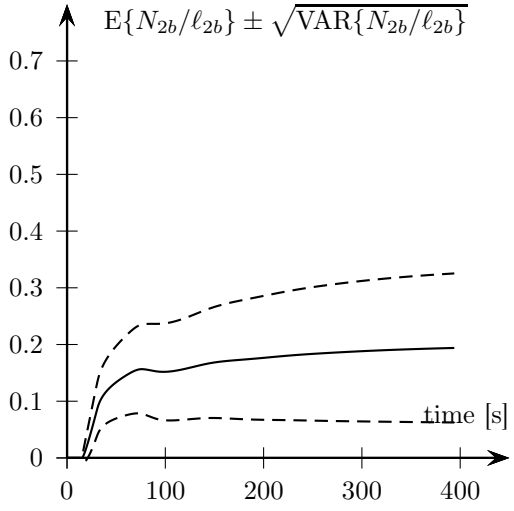
(a) Relative occupancy on link 1



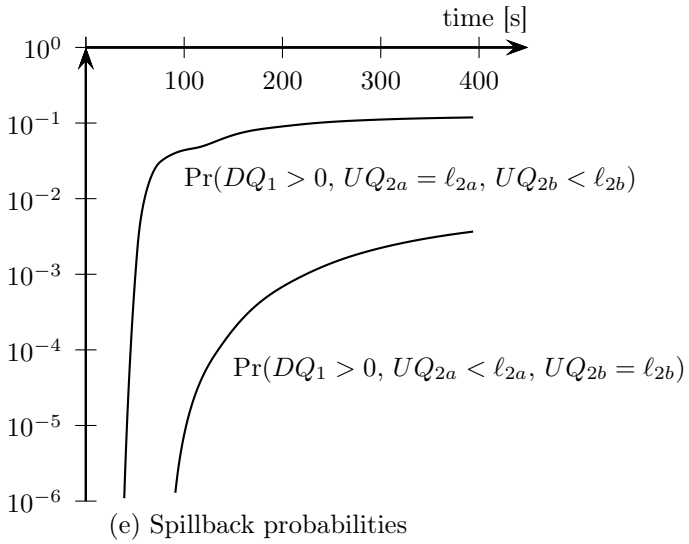
(b) Relative occupancy on link 3



(c) Relative occupancy on link 2a



(d) Relative occupancy on link 2b



(e) Spillback probabilities

Figure 13: Results in network experiment

When a given road network is mapped onto such a queueing network, every direction of a road and every intersection is mapped onto its own link respectively node subnetwork. Each link is represented by the four-queue system introduced by Osorio and Flötteröd (2015); this captures stochastic kinematic waves within the link as well as a joint distribution of the corresponding up- and downstream link boundary conditions. The node subnetworks comprise all queues defining the downstream boundary conditions of their ingoing links and all queues defining the upstream boundary conditions of their outgoing links.

The proposed model is validated in two stages. First, the accuracy of the analytical approximations at the queueing network level are validated versus simulation-based estimates. For this, a queueing network with complex dynamics that lead to multi-modal joint queue-length distributions is considered. A comparison in terms of transient expectations and standard deviations and all stationary bivariate queue-length distributions leads to the conclusion that the proposed model provides an accurate approximation of both the dynamics and the dependence structure. Second, the modeling of a road network is illustrated for a two-road tandem network and a more general network topology comprising a diverge and a merge node.

This modeling framework is operational and provides rich opportunities for future work. Five examples are given below.

Although the approximation model ensures mutual consistency of subnetwork distributions for their common queues, it does not guarantee the existence of an underlying joint distribution of which all subnetwork distributions are marginals. It is an open question if and how such consistency can be achieved. One may settle instead for an approximation error bound, which is yet to be established. Of more practical interest is the question of how to evaluate the network-wide dependencies captured by the model: Even if the proposed model approximates such a distribution, its computational advantage would be lost if an evaluation of this distribution would again require a complete state space enumeration.

The computational complexity of the proposed model scales linearly with the number of involved subnetworks. The state space of a single subnetwork comprises, however, still all possible state combinations of all queues contained in that subnetwork. For instance, the state space of a link subnetwork with space capacity ℓ is in the order of ℓ^3 (all four queues in the subnetwork have space capacity ℓ but are linearly dependent). The need to model long road segments or complex intersections with many in- and/or outgoing links motivates the further investigation of state space reduction techniques, such as the aggregation/disaggregation approach of Osorio and Yamani (Forthcoming).

The present article presents concrete linear, diverge and merge node specifications in order to demonstrate the SLTM's capability of modeling network traffic. These nodel models could be advanced by, for instance, the formulation of a general-topology node model (with an arbitrary number of in- and outgoing links) or the introduction of additional state variables that memorize the destination of individual vehicles queueing at or passing over the node.

In its present form, the model assumes transition rates to be exogenously given. In a network assignment context, where travelers choose routes and possibly departure times, turning and possibly also network arrival and departure rates become endogenous. Differently but related, a multi-commodity network assignment would require to model these rates per commodity. This relevant extension of the model could start out from Zhang et al. (2017), where a fixed and finite route choice set is considered, along with an analytical probabilistic route choice model, yet in a stationary setting. When considering dynamic network flows, Chabini (2001) provides an operational approach that iteratively attains consistency between link travel times and travel behavioral parameters. Another interesting, and yet to be explored, formulation would allow for en-route dynamic route choices.

The SLTM predicts the effect of stochasticity in network inflows, outflows, and between-link flow transitions. It does, in its present form, not predict the effect of stochasticity in, for instance, space capacities and speed limits (or, more general, wave speeds). Neither does the present article attempt to relate the stochastic SLTM model parameters to driving behavioral parameters, such as gap acceptance or reaction times. Further developing the SLTM in these directions would not only yield a richer model but also enable the development of measurement equations that would support the calibration of (stochastic) model parameters from real data.

Acknowledgments

The work of C. Osorio was partially supported by the National Science Foundation under Grant No. 1351512. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The constructive inquiries of three reviewers helped to broaden the scope of the article.

A Proofs

A.1 Proof of Proposition 1

Starting out from (7) with $(\mathbf{m},\mathbf{r},\mathbf{v}),(\mathbf{n},\mathbf{s},\mathbf{w}) \in \mathfrak{N}(\mathcal{W}) \times \mathfrak{N}(\partial\mathcal{W}) \times \mathfrak{N}(\mathcal{V}\backslash[\mathcal{W}\cup\partial\mathcal{W}])$, one has

$$\begin{aligned}
\frac{d}{d\tau}P(\mathbf{N}_{\mathcal{W}}=\mathbf{n}) &= \sum_{\mathbf{s},\mathbf{w}}\sum_{\mathbf{m},\mathbf{r},\mathbf{v}} t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{m},\mathbf{r},\mathbf{v}} P(\mathbf{N}=(\mathbf{m},\mathbf{r},\mathbf{v})) && (38)\\
&= \sum_{\mathbf{s},\mathbf{w}}\left[\sum_{(\mathbf{m},\mathbf{r},\mathbf{v})\neq(\mathbf{n},\mathbf{s},\mathbf{w})} t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{m},\mathbf{r},\mathbf{v}} P(\mathbf{N}=(\mathbf{m},\mathbf{r},\mathbf{v})) + t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{n},\mathbf{s},\mathbf{w}} P(\mathbf{N}=(\mathbf{n},\mathbf{s},\mathbf{w}))\right] && (39)\\
&= \sum_{\mathbf{s},\mathbf{w}}\sum_{(\mathbf{m},\mathbf{r},\mathbf{v})\neq(\mathbf{n},\mathbf{s},\mathbf{w})} \left[t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{m},\mathbf{r},\mathbf{v}} P(\mathbf{N}=(\mathbf{m},\mathbf{r},\mathbf{v})) - t^{\mathbf{m},\mathbf{r},\mathbf{v}}_{\mathbf{n},\mathbf{s},\mathbf{w}} P(\mathbf{N}=(\mathbf{n},\mathbf{s},\mathbf{w}))\right], && (40)
\end{aligned}$$

where the second term in the last row results from the definition of the main diagonal elements of a transition rate matrix. The addends in this expression are separated in two disjoint groups, a first group where $\mathbf{m}=\mathbf{n}$ and a second group where $\mathbf{m}\neq\mathbf{n}$. For the first group ($\mathbf{m}=\mathbf{n}$), one has

$$\begin{aligned}
&\sum_{\mathbf{s},\mathbf{w}}\sum_{(\mathbf{r},\mathbf{v})\neq(\mathbf{s},\mathbf{w})} \left[t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{n},\mathbf{r},\mathbf{v}} P(\mathbf{N}=(\mathbf{n},\mathbf{r},\mathbf{v})) - t^{\mathbf{n},\mathbf{r},\mathbf{v}}_{\mathbf{n},\mathbf{s},\mathbf{w}} P(\mathbf{N}=(\mathbf{n},\mathbf{s},\mathbf{w}))\right] && (41)\\
=\ &\sum_{\mathbf{s},\mathbf{w}}\left[\left(\sum_{\mathbf{r},\mathbf{v}} t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{n},\mathbf{r},\mathbf{v}} P(\mathbf{N}=(\mathbf{n},\mathbf{r},\mathbf{v})) - t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{n},\mathbf{s},\mathbf{w}} P(\mathbf{N}=(\mathbf{n},\mathbf{s},\mathbf{w}))\right)\ldots\right.\\
&\left.-\left(\sum_{\mathbf{r},\mathbf{v}} t^{\mathbf{n},\mathbf{r},\mathbf{v}}_{\mathbf{n},\mathbf{s},\mathbf{w}} P(\mathbf{N}=(\mathbf{n},\mathbf{s},\mathbf{w})) - t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{n},\mathbf{s},\mathbf{w}} P(\mathbf{N}=(\mathbf{n},\mathbf{s},\mathbf{w}))\right)\right] && (42)\\
=\ &\sum_{\mathbf{s},\mathbf{w}}\sum_{\mathbf{r},\mathbf{v}} \left[t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{n},\mathbf{r},\mathbf{v}} P(\mathbf{N}=(\mathbf{n},\mathbf{r},\mathbf{v})) - t^{\mathbf{n},\mathbf{r},\mathbf{v}}_{\mathbf{n},\mathbf{s},\mathbf{w}} P(\mathbf{N}=(\mathbf{n},\mathbf{s},\mathbf{w}))\right], && (43)
\end{aligned}$$

which is zero due to the symmetry of the double sum. Hence, only the second group with $\mathbf{m}\neq\mathbf{n}$ needs to be considered:

$$\frac{d}{d\tau}P(\mathbf{N}_{\mathcal{W}}=\mathbf{n}) = \sum_{\mathbf{s},\mathbf{w}}\sum_{\mathbf{m}\neq\mathbf{n}}\sum_{\mathbf{r},\mathbf{v}} \left[t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{m},\mathbf{r},\mathbf{v}} P(\mathbf{N}=(\mathbf{m},\mathbf{r},\mathbf{v})) - t^{\mathbf{m},\mathbf{r},\mathbf{v}}_{\mathbf{n},\mathbf{s},\mathbf{w}} P(\mathbf{N}=(\mathbf{n},\mathbf{s},\mathbf{w}))\right]. \quad (44)$$

Definition 3 ensures that (i) transition rates with $\mathbf{m}\neq\mathbf{n}$ (both in $\mathfrak{N}(\mathcal{W})$) and $\mathbf{v}\neq\mathbf{w}$ (both in $\mathfrak{N}(\mathcal{V}\backslash[\mathcal{W}\cup\partial\mathcal{W}])$) are zero and that (ii) nonzero transition rates with $\mathbf{m}\neq\mathbf{n}$ are independent of the concrete value of $\mathbf{v}=\mathbf{w}$. Accounting for this and inserting (9) yields

$$\begin{aligned}
\frac{d}{d\tau}P(\mathbf{N}_{\mathcal{W}}=\mathbf{n}) &= \sum_{\mathbf{s},\mathbf{w}}\sum_{\mathbf{m}\neq\mathbf{n}}\sum_{\mathbf{r}} \left[t^{\mathbf{n},\mathbf{s},\mathbf{w}}_{\mathbf{m},\mathbf{r},\mathbf{w}} P(\mathbf{N}=(\mathbf{m},\mathbf{r},\mathbf{w})) - t^{\mathbf{m},\mathbf{r},\mathbf{w}}_{\mathbf{n},\mathbf{s},\mathbf{w}} P(\mathbf{N}=(\mathbf{n},\mathbf{s},\mathbf{w}))\right] && (45)\\
&= \sum_{\mathbf{s},\mathbf{w}}\sum_{\mathbf{m}\neq\mathbf{n}}\sum_{\mathbf{r}} \left[t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{W}) P(\mathbf{N}=(\mathbf{m},\mathbf{r},\mathbf{w})) - t^{\mathbf{m},\mathbf{r}}_{\mathbf{n},\mathbf{s}}(\mathcal{W}) P(\mathbf{N}=(\mathbf{n},\mathbf{s},\mathbf{w}))\right] && (46)\\
&= \sum_{\mathbf{s}}\sum_{\mathbf{m}\neq\mathbf{n}}\sum_{\mathbf{r}} \left[t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{W}) P(\mathbf{N}_{\mathcal{W},\partial\mathcal{W}}=(\mathbf{m},\mathbf{r})) - t^{\mathbf{m},\mathbf{r}}_{\mathbf{n},\mathbf{s}}(\mathcal{W}) P(\mathbf{N}_{\mathcal{W},\partial\mathcal{W}}=(\mathbf{n},\mathbf{s}))\right]. && (47)
\end{aligned}$$

Substituting the main diagonal element of the local transition rate matrix defined in the second row of (9), one obtains

$$\frac{d}{d\tau}P(\mathbf{N}_{\mathcal{W}}=\mathbf{n}) = \sum_{\mathbf{s}}\left[\sum_{\mathbf{m}\neq\mathbf{n}}\sum_{\mathbf{r}} t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{W}) P(\mathbf{N}_{\mathcal{W},\partial\mathcal{W}}=(\mathbf{m},\mathbf{r})) + t^{\mathbf{n},\mathbf{s}}_{\mathbf{n},\mathbf{s}}(\mathcal{W}) P(\mathbf{N}_{\mathcal{W},\partial\mathcal{W}}=(\mathbf{n},\mathbf{s}))\right]. \quad (48)$$

Adding $\sum_{\mathbf{s}}\sum_{\mathbf{r}\neq\mathbf{s}} t^{\mathbf{n},\mathbf{s}}_{\mathbf{n},\mathbf{r}}(\mathcal{W}) P(\mathbf{n},\mathbf{r}) = 0$, where the third row of (9) ensures that all transition rates in this term are zero, yields

$$\begin{aligned}
\frac{d}{d\tau}P(\mathbf{N}_{\mathcal{W}}=\mathbf{n}) &= \sum_{\mathbf{s}}\left[\sum_{\mathbf{m}\neq\mathbf{n}}\sum_{\mathbf{r}} t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{W}) P(\mathbf{N}_{\mathcal{W},\partial\mathcal{W}}=(\mathbf{m},\mathbf{r})) + \sum_{\mathbf{r}} t^{\mathbf{n},\mathbf{s}}_{\mathbf{n},\mathbf{r}}(\mathcal{W}) P(\mathbf{N}_{\mathcal{W},\partial\mathcal{W}}=(\mathbf{n},\mathbf{r}))\right] && (49)\\
&= \sum_{\mathbf{s}}\sum_{\mathbf{m},\mathbf{r}} t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{W}) P(\mathbf{N}_{\mathcal{W},\partial\mathcal{W}}=(\mathbf{m},\mathbf{r})), && (50)
\end{aligned}$$

which coincides with (10). ■

A.2 Proof of Proposition 2

Consider the subnetwork $S \in \mathscr{S}(G)$ and cut out the region $G' = S \cup \partial S$ from the full network G. Note that $\mathcal{V}(G') = \mathcal{V}(S) \cup \partial\mathcal{V}(S)$. The subnetwork decomposition being triangle-free ensures that $\Psi_S(\mathbf{N}_{\partial\mathcal{V}(S)} \mid \mathbf{N}_{\mathcal{V}(S)})\Phi_S(\mathbf{N}_{\mathcal{V}(S)})$ is a probability distribution over $\mathfrak{N}(\mathcal{V}(S)) \times \mathfrak{N}(\partial\mathcal{V}(S))$. Definition 2 ensures that $\mathcal{W} = \mathcal{V}(S) \cap \mathcal{V}(T)$ overlaps with no subnetworks other than S and T, which implies $\partial\mathcal{W} = [\mathcal{V}(S) \cup \mathcal{V}(T)] \backslash \mathcal{W} \subset \mathcal{V}(S) \cup \partial\mathcal{V}(S)$. Letting $(\mathbf{m}, \mathbf{r}), (\mathbf{n}, \mathbf{s}) \in \mathfrak{N}(\mathcal{W}) \times \mathfrak{N}(\partial\mathcal{W})$, Proposition 1 hence allows to write

$$\begin{aligned} \frac{d}{d\tau}\Phi_S(\mathbf{N}_\mathcal{W} = \mathbf{n}) &= \sum_{\mathbf{s}} \sum_{\mathbf{m},\mathbf{r}} t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{W}) \Psi_S(\mathbf{N}_{\mathcal{V}(T)\backslash\mathcal{W}} = \mathbf{r}_{\mathcal{V}(T)\backslash\mathcal{W}} \mid \mathbf{N}_\mathcal{W} = \mathbf{m}) \ldots \\ &\quad \Phi_S(\mathbf{N}_{\mathcal{V}(S)} = (\mathbf{m}, \mathbf{r}_{\mathcal{V}(S)\backslash\mathcal{W}})). \end{aligned} \tag{51}$$

Substituting (14) leads to

$$\begin{aligned} \frac{d}{d\tau}\Phi_S(\mathbf{N}_\mathcal{W} = \mathbf{n}) &= \sum_{\mathbf{s}} \sum_{\mathbf{m},\mathbf{r}} t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{W}) \Phi_T(\mathbf{N}_{\mathcal{V}(T)\backslash\mathcal{W}} = \mathbf{r}_{\mathcal{V}(T)\backslash\mathcal{W}} \mid \mathbf{N}_\mathcal{W} = \mathbf{m}) \ldots \\ &\quad \Phi_S(\mathbf{N}_{\mathcal{V}(S)} = (\mathbf{m}, \mathbf{r}_{\mathcal{V}(S)\backslash\mathcal{W}})). \end{aligned} \tag{52}$$

Symmetric operations starting out from subnetwork T (cutting out $T \cup \partial T \subset G$, using Proposition 1 to express $\frac{d}{d\tau}\Phi_T(\mathbf{N}_\mathcal{W} = \mathbf{n})$) result in

$$\begin{aligned} \frac{d}{d\tau}\Phi_T(\mathbf{N}_\mathcal{W} = \mathbf{n}) &= \sum_{\mathbf{s}} \sum_{\mathbf{m},\mathbf{r}} t^{\mathbf{n},\mathbf{s}}_{\mathbf{m},\mathbf{r}}(\mathcal{W}) \Phi_S(\mathbf{N}_{\mathcal{V}(S)\backslash\mathcal{W}} = \mathbf{r}_{\mathcal{V}(S)\backslash\mathcal{W}} \mid \mathbf{N}_\mathcal{W} = \mathbf{m}) \ldots \\ &\quad \Phi_T(\mathbf{N}_{\mathcal{V}(T)} = (\mathbf{m}, \mathbf{r}_{\mathcal{V}(T)\backslash\mathcal{W}})). \end{aligned} \tag{53}$$

The equality of $\Phi_S(\mathbf{N}_\mathcal{W})$ and $\Phi_T(\mathbf{N}_\mathcal{W})$ implies that the right-hand sides of (52) and (53) are equal, which establishes the resulting equality of $\frac{d}{d\tau}\Phi_S(\mathbf{N}_\mathcal{W})$ and $\frac{d}{d\tau}\Phi_T(\mathbf{N}_\mathcal{W})$. ■

References

Akyildiz, I. F. and von Brand, H. (1994). Exact solutions to networks of queues with blocking-after-service, *Theoret. Comput. Sci.* **125**(1): 111–130.

Balakrishnan, V. (1997). *Schaum's Outline of Graph Theory*, 1st edn, McGraw-Hill.

Balsamo, S. and Donatiello, L. (1989). On the cycle time distribution in a two-stage cyclic network with blocking, *IEEE Trans. Software Eng.* **15**(10): 1206–1216.

Baskett, F., Mani Chandy, K., Muntz, R. and Palacios, F. (1975). Open, closed and mixed networks of queues with different classes of customers, *Journal of the ACM* **22**(2): 248–260.

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

Chabini, I. (2001). Analytical dynamic network loading problem, *Transportation Research Record* **1771**: 191–200.

Corthout, R., Flötteröd, G., Viti, F. and Tampere, C. (2012). Non-unique flows in macroscopic first-order intersection models, *Transportation Research Part B* **46**(3): 343–359.

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.

Deng, W., Lei, H. and Zhou, X. (2013). Traffic state estimation and uncertainty quantification based on heterogeneous data sources: A three detector approach, *Transportation Research Part B: Methodological* **57**: 132 – 157.

Flötteröd, G. and Rohde, J. (2011). Operational macroscopic modeling of complex urban intersections, *Transportation Research Part B* **45**(6): 903–922.

Grassmann, W. and Derkic, S. (2000). An analytical solution for a tandem queue with blocking, *Queueing Syst.* **36**(1-3): 221–235.

Griffiths, J. D., Leonenko, G. M. and Williams, J. E. (2008). Approximation to the transient solution of the M/Ek/1 queue, *INFORMS Journal on Computing* **20**(4): 510–515.

Gupta, S. (2011). A framework to span airport delay estimates using transient queuing models, *Technical report*, Massachusetts Institute of Technology.

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

Helbing, D. (2001). Traffic and related self-driven many-particle systems, *Rev. Mod. Phys.* **73**: 1067–1141.

Himpe, W., Corthout, R. and Tampere, M. C. (2016). An efficient iterative link transmission model, *Transportation Research Part B: Methodological* **92**: 170 – 190.

Jabari, S. and Liu, H. (2012). A stochastic model of traffic flow: theoretical foundations, *Transportation Research Part B* **46**(1): 156–174.

Jackson, J. R. (1957). Networks of waiting lines, *Oper. Res.* **5**(4): 518–521.

Jackson, J. R. (1963). Jobshop-like queuing systems, *Management Sci.* **10**(1): 131–142.

Kaczynski, W. H., Leemis, L. M. and Drew, J. H. (2012). Transient queueing analysis, *INFORMS Journal on Computing* **24**(1): 10–28.

Konheim, A. G. and Reiser, M. (1976). A queueing model with finite waiting room and blocking, *Journal of the Association for Computing Machinery* **23**(2): 328–341.

Konheim, A. G. and Reiser, M. (1978). Finite capacity queueing systems with applications in computer modeling, *SIAM J. Comput.* **7**(2): 210–229.

Kullback, S. and Leibler, R. (1951). On information and sufficiency, *Annals of Mathematical Statistics* **22**(1): 79–86.

Langaris, C. and Conolly, B. (1984). On the waiting time of a two-stage queueing system with blocking, *J. Appl. Probab.* **21**(3): 628–638.

Latouche, G. and Neuts, M. F. (1980). Efficient algorithmic solutions to exponential tandem queues with blocking, *SIAM Journal on Algebraic and Discrete Methods* **1**(1): 93–106.

Laval, J. A. and Chilukuri, B. R. (2014). The distribution of congestion on a class of stochastic kinematic wave models, *Transportation Science* **48**(2): 217–224.

Laval, J., He, Z. and Castrillon, F. (2012). Stochastic extension of newell's three-detector method, *Transportation Research Record: Journal of the Transportation Research Board* **2315**: 73–80.

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. and Khoshyaran, M. (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.

McCalla, C. and Whitt, W. (2002). A time-dependent queueing-network model to describe the life-cycle dynamics of private-line telecommunication services, *Telecommunication Systems* **19**(1): 9–38.

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

Nelson, P. and Kumar, N. (2006). Point constriction, interface, and boundary conditions for kinematic-wave model, *Transportation Research Record* **1965**: 60–69.

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

Odoni, A. R. and Roth, E. (1983). An empirical investigation og the transient behavior of stationary queueing systems, *Operations Research* **31**(3): 432–455.

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. (2009). 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 Flötteröd, G. (2015). Capturing dependency among link boundaries in a stochastic dynamic network loading model, *Transportation Science* **49**(2): 420–431.

Osorio, C., Flötteröd, G. and Bierlaire, M. (2011). Dynamic network loading: a stochastic differentiable model that derives link state distributions, *Transportation Research Part B* **45**(9): 1410–1423.

Osorio, C. and Wang, C. (2017). On the analytical approximation of joint aggregate queue-length distributions for traffic networks: a stationary finite capacity Markovian network approach, *Transportation Research Part B* **95**: 305–339.
Available at: http://web.mit.edu/osorioc/www/papers/osoWangAggDisagg.pdf .

Osorio, C. and Yamani, J. (Forthcoming). Analytical and scalable analysis of transient tandem Markovian finite capacity queueing networks, *Transportation Science* .
Available at: http://web.mit.edu/osorioc/www/papers/osoYamDynAggDisagg.pdf .

Peterson, M. D., Bertsimas, D. J. and Odoni, A. R. (1995a). Decomposition algorithms for analyzing transient phenomena in multiclass queueing networks in air transportation, *Operations Research* **43**(6): 995–1011.

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

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

Reibman, A. (1991). A splitting technique for Markov chain transient solution, *in* W. J. Stewart (ed.), *Numerical solution of Markov chains*, Marcel Dekker, Inc, New York, USA, chapter 19, pp. 373–400.

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

Schweitzer, P. (1991). A survey of aggregation-disaggregation in large Markov chains, *in* W. Stewart (ed.), *Numerical solutions of Markov chains*, Marcel Dekker Inc., pp. 63–88.

Sharma, O. P. and Gupta, U. C. (1982). Transient behavior of an M/M/1/N queue, *Stochastic Processes and their Applications* **13**: 327–331.

Sharma, O. P. and Shobha, B. (1988). Transient behaviour of a double-channel Markovian queue with limited waiting space, *Queueing Systems* **3**: 89–96.

Smits, E.-S., Bliemer, M. C., Pel, A. J. and van Arem, B. (2015). A family of macroscopic node models, *Transportation Research Part B: Methodological* **74**: 20 – 39.

Stewart, W. J. (1994). *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, NJ.

Stewart, W. J. (2000). Numerical methods for computing stationary distributions of finite irreducible Markov chains, *in* W. Grassmann (ed.), *Computational Probability*, Kluwer Academic Publishers, Boston, chapter 4.

Stewart, W. J. (2009). *Probability, Markov Chains, Queues, and Simulation*, Princeton University Press, Princeton, NJ.

Sumalee, A., Zhong, R. X., Pan, T. L. and Szeto, W. Y. (2011). Stochastic cell transmission model (SCTM): a stochastic dynamic traffic model for traffic state surveillance and assignment, *Transportation Research Part B* **45**(3): 507–533.

Tampere, C., Corthout, R., Cattrysse, D. and Immers, L. (2011). A generic class of first order node models for dynamic macroscopic simulations of traffic flows, *Transportation Research Part B* **45**(1): 289–309.

Tampere, C., van Arem, B. and Hoogendoorn, S. (2003). Gas-kinetic flow modeling including continuous driver behavior models, *Transportation Research Record* **1852**: 231–238.

Whitt, W. (1999). Decomposition approximations for time-dependent Markovian queueing networks, *Operations Research Letters* **24**: 97–103.

Yperman, I., Logghe, S., Tampere, C. and Immers, B. (2006). The multi-commodity link transmission model for dynamic network loading, *Proceedings of the 85. Annual Meeting of the Transportation Research Board*, Washington, DC, USA.

Zhang, C., Osorio, C. and Flötteröd, G. (2017). Efficient calibration techniques for large-scale traffic simulators, *Transportation Research Part B* **97**: 214–239.

Zhong, R. X., Sumalee, A., Pan, T. L. and Lam, W. H. (2013). Stochastic cell transmission model for traffic network with demand and supply uncertainties, *Transportmetrica A: Transport Science* **9**(7): 567–602.