Transportation Research Part B 95 (2017) 305–339

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

On the analytical approximation of joint aggregate queue-length distributions for traffic networks: A stationary finite capacity Markovian network approach

Carolina Osorio*, Carter Wang

Massachusetts Institute of Technology (MIT), Department of Civil & Environmental Engineering, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 19 December 2015
Revised 22 July 2016
Accepted 25 July 2016
Available online 15 November 2016

ABSTRACT

This paper is motivated by recent results in the design of signal plans for Manhattan that highlight the importance of providing signal control algorithms with an analytical description of between-link dependencies. This is particularly important for congested networks prone to the occurrence of spillbacks. This paper formulates a probabilistic network model that proposes an aggregate description of the queue-length, and then approximates the joint aggregate queue-length distribution of subnetworks. The goal is to model between-queue dependencies beyond first-order moments, yet to do so in a tractable manner such that these techniques can be used for optimization purposes.

This paper models an urban road network as a finite space capacity Markovian queueing network. Exact evaluation of the stationary joint queue-length distribution of such a network with arbitrary size and topology can be obtained numerically. Nonetheless, the main challenge to such an approach remains the dimensionality of the network state space, which is exponential in the number of queues. This paper proposes to address the dimensionality issue by: 1) describing the state of the network aggregately, and 2) decomposing the network into overlapping subnetworks. We propose an analytical approximation of the stationary aggregate joint queue-length distribution of a subnetwork. The model consists of a system of nonlinear equations with a dimension that is linear, instead of exponential, in the number of queues and that is independent of the space capacity of the individual queues. The method is derived for tandem Markovian finite capacity queueing networks. The proposed model is computationally tractable and scalable, it can be efficiently used for the higher-order distributional analysis of large-scale networks. The model is validated versus simulation estimates and versus other decomposition methods. We then use it to address an urban traffic control problem. We show the added value of accounting for higher-order spatial between-queue dependency information in the control of congested urban networks.

1. Introduction

Current challenges faced, and goals set, by major urban transportation agencies call for the development of traffic models that can provide a probabilistic, yet also computationally tractable, description of traffic. A few examples of such challenges

* Corresponding author.
E-mail addresses: osorioc@mit.edu (C. Osorio), carterw@mit.edu (C. Wang).

http://dx.doi.org/10.1016/j.trb.2016.07.013
0191-2615/

Table 1
Properties of the queue-length distributions derived by our past queueing-theoretic traffic models.

Model	Marginal	Joint	Stationary	Transient	Disaggregate	Aggregate
Osorio (2010, Chap. 4)	✓		✓		✓	
Osorio and Chong (2015)	✓		✓		✓	
Osorio et al. (2011)	✓			✓	✓	
Osorio and Flötteröd (2015)	✓			✓	✓	
Chong and Osorio (2016)	✓			✓	✓	
Osorio and Yamani (2016)		✓		✓		✓
Osorio and Wong (this paper)		✓	✓			✓

and goals follow. First, various major agencies have identified the improvement of network reliability and network robustness as a primary and critical goal (Department of Transportation, 2008; Texas Transportation Institute, 2012; Transport for London, 2010). To address reliable and robust formulations of traditional urban transportation optimization problems there is a need for probabilistic and tractable traffic models. Second, for congested networks with short links and grid topologies, forecasting and mitigating the occurrence of vehicular spillbacks is a major challenge. For instance, signal control work within Manhattan, that we carried out in collaboration with the New York City Department of Transportation, has shown that for congested urban networks with intricate traffic patterns, it is important to provide the algorithms with an analytical description of between-link dependencies (Osorio et al., 2014). The formulation of probabilistic traffic models that describe the spatial between-link dependencies, and hence the spatial propagation of congestion (e.g., vehicular spillbacks, gridlocks) is needed to inform the operations of such networks. Third, the estimation of network-wide or path performance measures involves the approximation of joint network or joint path distributions. This calls for a description of between-link dependencies that goes beyond first-order moments. In other word, it goes beyond the use of the expectation of the performance measure to consider its variance or even its full distribution.

The model proposed in this paper contributes to address the challenges mentioned above. It is a probabilistic, analytical and tractable, traffic model that approximates the joint aggregate queue-length distribution of adjacent links. Hence, it yields a detailed and higher-order description of between-link dependencies.

The focus of this paper is on queueing network theory models. There have been numerous attempts to develop probabilistic link models (i.e., models of vehicular traffic along a homogenous road segment) based on queueing theory. Past work includes Heidemann (1994); 1996); 2001); Heidemann and Wegmann (1997); Jain and Smith (1997); Tanner (1962); Van Woensel and Vandaele (2007). For a recent review, see Osorio (2010, Chap 4.2). More recently, a stochastic formulation of the link-transmission model (Yperman, 2007) was proposed (Osorio and Flötteröd, 2015; Osorio et al., 2011). The link transmission model is an operational instance of Newell's simplified theory of kinematic waves (Newell, 1993).

Existing probabilistic urban traffic models approximate either first-order moments of the main link performance metrics or the marginal probability distributions of these link metrics. This paper proposes an approach to efficiently approximate the joint queue-length distribution of adjacent links. The purpose is to account for between-link dependencies beyond first-order moments, yet to do so in a tractable manner such that these techniques can be used for optimization purposes. Such higher-order distributional information can then lead to more accurate path and network-wide performance estimates and ultimately enhanced network design and traffic management strategies. The main challenges that arise in such an approach are the dimensionality of the state space and the complexity of modeling network-wide dependency both analytically and tractably.

This paper models an urban road network as a finite capacity Markovian queueing network. Table 1 summarizes our past work in the formulation and use of finite capacity queueing theory for road traffic. Each model that appears in the table yields queue-length distributions. The table indicates whether the distributions are: (i) marginal or joint, (ii) stationary or transient, (iii) aggregate or disaggregate. The last row considers the model formulated in this paper. This table indicates that most past work has focused on the approximation of marginal distributions. The work of Osorio and Yamani (2016) yields joint distributions. It provides a transient formulation, which comes at a higher computational cost than the stationary joint model presented in the current paper, which is formulated as a nonlinear system of equations, and is hence both scalable (i.e., relevant for the analysis of large-scale networks) and computationally efficient to evaluate. The use of Markovian finite capacity queueing theory models for vehicular traffic has followed two directions. First, for uninterrupted traffic, Markovian link models that are consistent with the traditional deterministic traffic flow theory of kinematic waves of Lighthill and Witham (1955); Richards (1956) have been proposed (Osorio and Flötteröd, 2015; Osorio et al., 2011). Such models provide a detailed description of uninterrupted traffic. Nonetheless, their formulation for full networks has yet to be explored. Ideas from the proposed model can be combined with such link models in order to enhance their scalability and tractability. More specifically, the proposed model can serve to track in a scalable way between-link boundary conditions, while the detailed link models mentioned above can serve to track within-link traffic conditions.

Second, for interrupted traffic (e.g., signal-controlled traffic), Markovian models that accurately describe congested traffic conditions have not been proposed. Nonetheless, tractable Markovian models have been successfully combined with accurate, yet intractable, simulation-based models of interrupted traffic (e.g., stochastic microscopic traffic simulators) in order to enhance the computational efficiency of simulation-based optimization (SO) algorithms. The analytical models embedded

within the SO algorithms are based on various extensions of the model of Osorio (2010, Chap. 4). These algorithms have been used for various traffic management and model calibration case studies, including Lausanne (Switzerland), New York City (USA) and ongoing work in San Diego (USA) and Berlin (Germany) (Osorio and Bierlaire, 2013; Osorio et al., 2014; Osorio and Chong, 2015; Osorio and Nanduri, 2015a; 2015b; Osorio et al., 2016). Since the proposed model is an analytical, differentiable and tractable model, it can also be used to perform efficient SO. Past SO work has used Markovian models that approximate the marginal queue-length distributions. The proposed model approximates joint distributions, and hence can be used to address SO problems that require a higher-order description of between-link dependencies.

Markovian finite capacity queueing network (FCQN) models are relevant for the optimization of urban traffic for two main reasons. First, each link of an urban network is modeled as one (or multiple) finite capacity queue(s). The term capacity refers to the space capacity of the queue. Such models differ from traditional (infinite capacity) queueing models in that they assume there is an upper bound on the number of vehicles that can fit within a queue. Similarly, they impose an upper-bound on the inflow to each link. Hence, high flows, which are consistent with the Poisson distribution's fat right tail, cannot enter the link.

Second, the models account for the occurrence and impact of vehicular spillbacks. The complexity of describing spillback phenomena in analytical, let alone differentiable, form has limited the formulation of finite space capacity models. Our past work, and the present paper, provide an analytical approximation of spillback probabilities and of the impact of spillbacks on upstream link flow capacities. Recent work has highlighted the importance of providing transportation optimization algorithms with an analytical description of between-link dependencies, and in particular, of spillback phenomena (Osorio et al., 2014).

These Markovian models are based on strong distributional assumptions (homogeneous or nonhomogeneous Poisson arrival patterns, exponential inter-departure times). Our past work has overcome the main limitations of these distributional assumptions by formulating the parameters of these distributions as endogenous model variables that account for between-link dependencies (e.g., spillbacks).

The analytical stationary analysis of FCQNs is complex for various reasons. Firstly, unlike Jackson networks or BCMP networks (Baskett et al., 1975; Jackson, 1957; 1963), such models do not have product-form joint queue-length distributions, i.e. their joint distribution cannot be decomposed as a product of its marginals. Secondly, finite capacity leads to potential spillbacks (referred to in queueing theory as blocking). That is, the queue of vehicles along a road may extend beyond the road to upstream roads. Analyzing the blocking phenomenon analytically is challenging, as illustrated in Osorio and Bierlaire (2009). Blocking is not captured with infinite capacity queues, but is prevalent in a variety of real-world congested networks, ranging from hospital networks (Osorio and Bierlaire, 2009) to biological protein synthesis networks (Osorio and Bierlaire, 2012).

This paper models an urban traffic network as a finite capacity Markovian queueing network. As in our past urban traffic work, each queue is an M/M/1/k queue (Osorio, 2010; Osorio and Flötteröd, 2015; Osorio et al., 2011). Exact numerical evaluation of the stationary joint distribution of a Markovian network with arbitrary size and topology can be obtained by solving the global balance equations (presented in Section 2.1). A detailed description of these numerical methods can be found in Stewart (2000). Nonetheless, the main challenge to such an approach remains the dimensionality of the network state space. For instance, for a network with m queues each with space capacity k (hereafter referred to as capacity), the dimension of the state space of the joint distribution is $(k+1)^m$. The dimension is exponential in the number of queues. For realistically-sized networks, exact numerical techniques lack computational tractability. Thus, most analytical analysis of FCQNs consist of approximation methods.

The most popular approximate approaches are decomposition methods, which reduce the dimensionality of the system under study by decomposing the network into smaller subnetworks (also called subsystems). Each subnetwork is then modeled individually. Some dependency is captured by approximating the structural parameters of a subnetwork (e.g. arrival rates, service rates) as a function of the performance of other subnetworks (e.g., flow conservation equations). Most of the research in this field has been pioneered and driven by the manufacturing field, as illustrated by the review of Dallery and Gershwin (1992). There is extensive research in decomposition methods that decompose the network into single queues and approximate the marginal distributions of each queue (e.g., Hillier and Boling, 1967; Takahashi et al., 1980; Altiok, 1982; Gershwin, 1987; Altiok and Perros, 1987; Kerbache and Smith, 1987; 1988; Jun and Perros, 1989; 1990; Cheah and Smith, 1994; Singh and Smith, 1997; Tolio and Gershwin, 1998; Tahilramani et al., 1999; Gershwin and Burman, 2000; Kerbache and Smith, 2000; Korporaal et al., 2000; Gupta and Kavusturucu, 2000; Koizumi et al., 2005; Osorio and Bierlaire, 2009). Decomposition methods that consider subnetworks with two or three queues have also been proposed (Alfa and Liu, 2004; Brandwajn and Jow, 1985; 1988; Schmidt and Jackman, 2000; van Vuuren et al., 2005).

Another approach to partially address the curse of dimensionality is to derive an aggregate distribution that is consistent with the underlying disaggregate distribution. Aggregation-disaggregation techniques for queueing networks have addressed this issue in the past. Schweitzer (1991) presents a survey of aggregation-disaggregation techniques. To the best of our knowledge, the first such approach is that of Takahashi (1975). The aggregation-disaggregation process is referred to as lumping. It considers an arbitrarily high-dimensional Markov chain. It clusters the states of the chain into a set of aggregate states. The high-dimensional global balance equations that define the disaggregate joint distribution are solved by an exact numerical technique that iteratively solves two lower-dimensional systems of linear equations which are the global balance equations for: 1) the aggregate joint distribution; 2) the local disaggregate distribution of a given aggregate state. Schweitzer (1984) formulates the Takahashi (1975) approach for arbitrary topology and size Markovian FCQNs, and for both the station-

ary and transient distributions. The method of Schweitzer (1984) yields an exact numerical evaluation of the aggregate joint distribution. The state space of the aggregate joint network distribution is still exponential in the number of queues, hence this method does not fully overcome the dimensionality issue. Approximate aggregation-disaggregation methods suitable for small-size Markovian FCQN networks have been proposed by Takahashi (1985) and by Song and Takahashi (1991).

To summarize, decomposition methods typically have a model complexity that is linear in the number of queues yet depends on the space capacity of the queues. In an urban network these space capacities can be large, and hence a decomposition approach may lack tractability. On the other hand, aggregation-disaggregation methods typically have a model complexity that is exponential in the number of queues yet is independent of the space capacity of the queues. In an urban network with a large number of queues, a traditional aggregation-disaggregation approach may not be sufficiently tractable.

This paper proposes to address the dimensionality issue by combining ideas from both decomposition methods and aggregation-disaggregation methods. In other words, we propose to: 1) decompose the network into (lower-dimensional) overlapping subnetworks of queues, and 2) describe the state of each queue aggregately. This leads to a model complexity that is both linear in the number of queues (just like decomposition methods) and is also independent of the space capacity of the individual queues (just like aggregation-disaggregation techniques). This combination of ideas leads to a tractable model that can be efficiently used for the higher-order distributional analysis of large-scale networks.

The proposed model yields an analytical approximation of the stationary aggregate joint queue-length distribution of a subnetwork. The model is formulated as a system of nonlinear equations with a dimension that is linear in the number of queues and that is independent of the queue space capacities. The method is derived for tandem Markovian FCQNs. The approach considers a stationary regime and combines ideas from the methods of Takahashi (1975); 1985) and Schweitzer (1984) along with ideas from other decomposition techniques for FCQNs (Osorio and Bierlaire, 2009) and from probabilistic road traffic models (Osorio, 2010, Chap. 4).

The proposed methodology is presented in Section 2. It is validated versus simulation results and versus other decomposition methods (Section 3). It is then used to investigate the added value of accounting for full distributional spatial dependency in the context of urban traffic signal control (Section 4). Section 5 presents the main conclusions.

2. Model formulation

Section 2.1 presents the aggregation technique of Schweitzer (1984) for a general Markov chain. We formulate this aggregation technique for a single finite space capacity queue (Section 2.2) and use this formulation to propose an aggregation technique for an isolated three queue tandem network (Section 2.3). Section 2.4 generalizes the technique for a tandem network with an arbitrary number of queues.

2.1. Aggregation-disaggregation framework

In order to address the dimensionality issues mentioned in the previous section, we use the aggregation technique described by Schweitzer (1984). This section presents its main ideas. The technique considers a continuous or discrete time Markov chain with a finite and large state space. The Markov chain is assumed aperiodic and communicative. Let Ω denote the state space with $\text{card}(\Omega) = M$. The probability of being in an individual state $i \in \Omega$ at steady state is denoted by π_i. The rate at which a transition from state i to j, $i \neq j$, $(i, j) \in \Omega^2$, can take place is given by q_{ij}. The steady state probabilities satisfy:

$$\begin{cases} \pi_i \sum_{j \in \Omega \setminus i} q_{ij} = \sum_{j \in \Omega \setminus i} \pi_j q_{ji}, & \forall i \in \Omega \quad \text{(a)} \\ \sum_{i \in \Omega} \pi_i = 1. & \text{(b)} \end{cases} \tag{1}$$

The above system of equations is referred to as the global balance equations. For a detailed derivation, see for instance Chapter 4.5 in Larson and Odoni (1981).

The global balance equations can be rewritten in matrix format as:

$$\begin{cases} \pi Q = 0 & \text{(a)} \\ \sum_{i \in \Omega} \pi_i = 1, & \text{(b)} \end{cases} \tag{2}$$

where Q is known as the transition rate matrix, and is defined as:

$$Q_{ij} = \begin{cases} q_{ij}, & \text{if} \quad i \neq j, \\ -\sum_{k \in \Omega \setminus i} q_{ik} & \text{if} \quad i = j. \end{cases} \tag{3}$$

For Markov chains with a large number of states, Schweitzer (1984) proposes to partition the M states into $\overline{M}$ aggregate disjoint states, such that $\overline{M} \ll M$. Let $\bar{\Omega}$ denote the set of aggregate states.

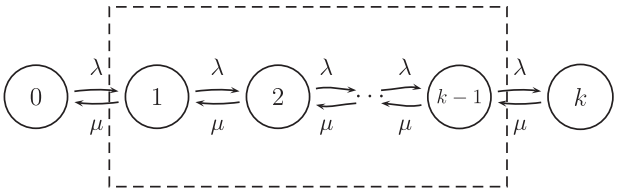

Fig. 1. State transition diagram for a single disaggregate queue.

Let Ω_a denote the set of disaggregate states within aggregate state a. The probability of being in aggregate state a, $\bar{\pi}_a$, is defined as:

$$\bar{\pi}_a = \sum_{i \in \Omega_a} \pi_i. \tag{4}$$

The global balance equations for the aggregate distribution are given by:

$$\begin{cases} \bar{\pi}_a \sum\limits_{b \in \bar{\Omega} \setminus a} \bar{q}_{ab} = \sum\limits_{b \in \bar{\Omega} \setminus a} \bar{\pi}_b \bar{q}_{ba}, \quad \forall a \in \bar{\Omega} \quad \text{(a)} \\ \sum\limits_{a \in \bar{\Omega}} \bar{\pi}_a = 1. \text{(b)} \end{cases} \tag{5}$$

where, $\bar{q}_{ab}$ is the transition rate from aggregate state a to aggregate state b, and is referred to as an aggregate transition rate.

Schweitzer (1984) relates the aggregate transition rates to the disaggregate transition rates made from all disaggregate states into aggregate state a through the following equation (which corresponds to Equation (2.9) of Schweitzer (1984)):

$$\bar{q}_{ab} = \frac{\sum_{j \in \Omega_a} \sum_{i \in \Omega_b} \pi_j q_{ji}}{\sum_{k \in \Omega_a} \pi_k}, \quad (b, a) \in \bar{\Omega}^2, b \neq a. \tag{6}$$

This equation implies that the aggregate transition rates can be computed in exact form if both the disaggregate distribution (π) and the disaggregate transition rates (q) are known. Typically, q is known but π is unknown. This equation can be rewritten as:

$$\bar{q}_{ab} = \sum_{j \in \Omega_a} \sum_{i \in \Omega_b} \left(\frac{\pi_j}{\sum_{k \in \Omega_a} \pi_k} \right) q_{ji}, \quad (b, a) \in \bar{\Omega}^2, b \neq a. \tag{7}$$

The term in parenthesis represents the conditional probability of being in disaggregate state j conditional on being in aggregate state a. We refer to this conditional probability as a disaggregation probability. Eq. (7) indicates that the only disaggregation probabilities needed to compute $\bar{q}_{ab}$ are those where $q_{ji} > 0$. In other words, we need only to consider disaggregate states j that allow for transitions into the aggregate state b, knowledge of the full disaggregation distribution π is not needed.

In the next section, we derive the exact form of Eq. (7) for an M/M/1/k system (cf. Eq. (13)). This allows us to identify the set of disaggregation probabilities needed to obtain the aggregate rates $\bar{q}$. We then use these insights to formulate an approximate expression for the aggregate transition rates of a tandem network with three queues. The latter is the main building block of the proposed network model.

2.2. Aggregate description of a single queue

In this section, we consider a single M/M/1/k queue and derive expressions for its aggregate transition rates. These expressions are then used in Sections 2.3 and 2.4 to derive the aggregate model for a tandem queueing network. The state of a queue is described by the number of jobs (e.g., vehicles), N, in the queueing system. The state space is given by $\Omega = \{0, 1, .., k\}$, where $k \in \mathbb{Z}^+$ is the space capacity. The corresponding state transition diagram is displayed in Fig. 1. Each circle denotes a state. The arrows denote possible transitions between the states and their corresponding rates. In this case, arrivals are determined by the arrival rate, $\lambda \geq 0$, and departures are determined by the service rate, $\mu > 0$.

We aggregate the $k+1$ states into the following three states: the queue is empty, the queue is full, the queue is neither empty nor full. The choice of these three states is based on insights from urban traffic intersection (i.e., node) models, where between-link interactions (i.e. interactions of links that are connected via an intersection) are mainly determined based on whether a vehicle is ready to be sent from an upstream link to a downstream link (i.e. non-empty upstream link) and whether there is space downstream to receive this vehicle (i.e. non-full downstream link). There are now three aggregate states: state 0, state k, and the state defined by the dashed lines in Fig. 1.

Fig. 2 represents the state transition diagram of the aggregate queueing system. The states 0, 1 and 2 denote, respectively, the disaggregate states 0, $\{1, \ldots, k-1\}$ and k. As represented in Fig. 2, the aggregate system is now fully described by a set

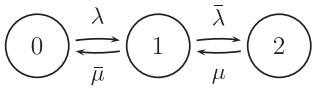

Fig. 2. State transition diagram for a single aggregate queue.

of four rates: λ, μ, $\bar{\lambda}$, and $\bar{\mu}$, where $\bar{\mu}$ and $\bar{\lambda}$ describe the transition rates from the new aggregate state to one of the other states (they are the aggregate transition rates).

The aggregate states are described by the random variable N_A:

- Empty queue: $N_A = 0, \quad \Omega_0 = \{N = 0\}$,
- Non-empty and non-full queue: $N_A = 1, \quad \Omega_1 = \{N \in [1, k-1]\}$,
- Full queue: $N_A = 2, \quad \Omega_2 = \{N = k\}$.

Hereafter the disaggregate (resp. aggregate) state probability π_i (resp. $\bar{\pi}_i$) is denoted $P(N = i)$ (resp. $P(N_A = i)$). The global balance equations satisfied by the aggregate state probabilities are:

$$\begin{cases} \lambda P(N_A = 0) = \bar{\mu} P(N_A = 1) & \text{(a)} \\ \mu P(N_A = 2) = \bar{\lambda} P(N_A = 1) & \text{(b)} \\ \sum_{i=0}^{2} P(N_A = i) = 1 & \text{(c)} \end{cases} \tag{8}$$

Following Eq. (6), the aggregate transition rates are given by:

$$\begin{cases} \bar{\lambda} = \dfrac{\sum_{j\in\Omega_1} \sum_{i\in\Omega_2} P(N = j) q_{ji}}{\sum_{j\in\Omega_1} P(N = j)} & \text{(a)} \\ \bar{\mu} = \dfrac{\sum_{j\in\Omega_1} \sum_{i\in\Omega_0} P(N = j) q_{ji}}{\sum_{j\in\Omega_1} P(N = j)}. & \text{(b)} \end{cases} \tag{9}$$

In an M/M/1/k queue we have (see Fig. 1):

$$q_{jk} = \begin{cases} \lambda & \text{if } j = k-1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$q_{j0} = \begin{cases} \mu & \text{if } j = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Inserting Eqs. (10) and (11) into (9), and noting that $\Omega_0 = \{0\}$ and $\Omega_2 = \{k\}$, we obtain:

$$\begin{cases} \bar{\lambda} = \dfrac{P(N = k-1)\lambda}{\sum_{j\in\Omega_1} P(N = j)} = \lambda \dfrac{P(N = k-1)}{P(N_A = 1)} & \text{(a)} \\ \bar{\mu} = \dfrac{P(N = 1)\mu}{\sum_{j\in\Omega_1} P(N = j)} = \mu \dfrac{P(N = 1)}{P(N_A = 1)}. & \text{(b)} \end{cases} \tag{12}$$

The above equations can be rewritten as:

$$\begin{cases} \bar{\lambda} = \lambda\, P(N = k-1 \mid N_A = 1) & \text{(a)} \\ \bar{\mu} = \mu\, P(N = 1 \mid N_A = 1). & \text{(b)} \end{cases} \tag{13}$$

In summary, in order to determine the aggregate transition rates ($\bar{\lambda}$ and $\bar{\mu}$), we need to evaluate the probabilities $P(N = k-1 \mid N_A = 1)$ and $P(N = 1 \mid N_A = 1)$. We call these probabilities disaggregation probabilities, since they represent the probabilities of being in a disaggregate state of a given aggregate state.

For a single M/M/1/k queue, there is a closed-form expression for the stationary queue-length distribution (see, e.g., Bocharov et al., 2004):

$$P(N = n) = \frac{(1-\rho)\rho^n}{1-\rho^{k+1}} \quad \forall n \in [0, k], \tag{14}$$

where ρ is known as the traffic intensity and is defined as λ/μ. This expression holds for $\rho \neq 1$. We assume hereafter that $\rho \neq 1$. Thus, we can obtain an exact closed-form expression for the disaggregation probabilities.

$$P(N=1 \mid N_A=1) = \frac{P(N=1)}{P(N_A=1)} = \left(\frac{(1-\rho)\rho}{1-\rho^{k+1}}\right)\left(\sum_{j=1}^{k-1}\frac{(1-\rho)\rho^j}{1-\rho^{k+1}}\right)^{-1}$$
$$= \rho\left(\sum_{j=1}^{k-1}\rho^j\right)^{-1} = \rho\left(\rho\frac{1-\rho^{k-1}}{1-\rho}\right)^{-1} \tag{15}$$

$$P(N=1 \mid N_A=1) = \frac{1-\rho}{1-\rho^{k-1}}. \tag{16}$$

We can proceed similarly to obtain:

$$P(N=k-1 \mid N_A=1) = \frac{(1-\rho)\rho^{k-2}}{1-\rho^{k-1}}. \tag{17}$$

2.3. Aggregate description of a three queue tandem network

We consider a network of three queues in a tandem (i.e. series, also called linear) topology. This is the simplest topology in which a queue is affected by both upstream and downstream traffic conditions. Hereafter index i refers to a queue index.

Queue i ($i \in \{1, 2, 3\}$) has finite space capacity $k_i \in \mathbb{Z}^+$ and independent exponentially distributed service times with parameter μ_i. For each queue, external arrivals (i.e. arrivals that come from outside the network) follow a Poisson process with rate parameter γ_i. The joint aggregate state probabilities are denoted by $P(N_A = s)$, where an aggregate state s is defined as the triplet: $s = (j_i, j_{i+1}, j_{i+2})$, and $j_i \in \{0, 1, 2\}$. For a three queue network with three aggregate states, the state space is aggregated into $3^3 = 27$ distinct states. The dimension of the state space is now independent of the individual queue capacities.

The main challenges, when extending the approach from a single queue to a network of queues, arise as a result of the possibility of blocking. Blocking occurs when a job (e.g., a vehicle) completes service yet cannot proceed downstream because the downstream queue is full. This is known as spillback in urban traffic. We assume here blocking-after-service (Balsamo et al., 2001), where a blocked job continues to occupy the underlying server until it is unblocked. Hence, it prevents other jobs in the queue from receiving service. Blocking-after-service is also known as manufacturing blocking, production blocking, type 1 blocking or transfer blocking. Thus, a job that cannot proceed downstream due to lack of space is blocked, while also blocking the use of the underlying server. A job becomes unblocked when its downstream destination queue can accommodate it.

Two main challenges that arise due to blocking are:

1. the rate of job departures at a blocked queue depends on both the states and the service rates of downstream queues,
2. a service completion at a blocking queue (i.e. a queue that is blocking jobs at upstream queues) triggers instantaneous state changes at upstream blocked queues.

Methods to approximate stationary marginal queue-length distributions that explicitly describe blocking states have been proposed (e.g., Osorio and Bierlaire (2009)). They illustrate the complexity of approximating the blocking probabilities and the effective service rates (i.e., the service rates that account for the occurrence of blocking). In this paper, we propose simple approximations to account for both challenges, these are described in Sections 2.3.1 and 2.3.2.

2.3.1. Aggregate transition rates

Within a three queue network, we assume that the aggregate transition rates for a given queue i are given by (13). These depend on the disaggregation probabilities. We assume the latter to have the functional form given by Eqs. (16) and (17). Hence, they depend on the traffic intensity ρ of the underlying queue, which is unknown. In this section, we first present the approximation of ρ, followed by the approximation of the disaggregation probabilities.

Traffic intensity

The traffic intensity is defined as the ratio between an arrival rate and a service rate. In a finite capacity queueing network, the prevailing arrival and service rates of a given queue may be state-dependent. This is due to the occurrence of blocking. To illustrate this, consider a given queue i that is the most upstream queue of a three-queue tandem network. Consider a job that is occupying a server at queue i. It can either be:

1. undergoing service, which will be completed with rate μ_i,
2. blocked by its directly downstream queue which is itself not blocked, the job will therefore be unblocked with rate μ_{i+1},
3. blocked by its directly downstream queue which is itself blocked, the job will therefore be unblocked with rate μ_{i+2}.

Thus, depending on the state of that job (under service, blocked and if so blocked by which queue), the queue will have a different prevailing service rate.

For a given queue i and a given aggregate state s, let $\rho_{i,s}$, λ_i and μ_i, s denote, respectively, the traffic intensity, the arrival rate and the service rate. The traffic intensity is given by:

$$\rho_{i,s} = \frac{\lambda_i}{\mu_{i,s}}. \tag{18}$$

This equation states that we use a queue-dependent but state-independent arrival rate, along with a queue- and state-dependent service rate. Let us know explain how we approximate each of these rates.

- **Service rate.** For state $s = (1, j_{i+1}, j_{i+2})$, the state-dependent (prevailing) service rate, $\mu_{i,s}$, is given by:

$$\mu_{i,s} = \begin{cases} \mu_i & \text{if} \quad j_{i+1} < 2 \\ \mu_{i+1} & \text{if} \quad j_{i+1} = 2 \quad \text{and} \quad j_{i+2} < 2 \\ \mu_{i+2} & \text{if} \quad j_{i+1} = 2 \quad \text{and} \quad j_{i+2} = 2. \end{cases} \tag{19}$$

  This approximation states that if queue i has one or several consecutive downstream queues that are full, then its prevailing service rate is that of the most downstream queue that is full. Recall that for queue i, the only aggregate state with more than one disaggregate state (i.e. the only state where disaggregation probabilities are needed) is aggregate state 1. This is why, when approximating the disaggregation probabilities of queue i, we need only to consider states s with $j_i = 1$.
- **Arrival rate.** The arrival rate of queue i is obtained by solving the flow conservation equation (which is derived in Osorio and Bierlaire (2009, Eq. (3)), and is rewritten with the same notation as in this paper in Osorio and Chong (2015, Eq.(3a))):

$$\lambda_i = \gamma_i + \frac{\lambda_{i-1} P(N_{i-1} < k_{i-1})}{P(N_i < k_i)}, \tag{20}$$

  where:

$$P(N_i < k_i) = P(N_{A,i} < 2) \tag{21}$$

  and N_i (resp. $N_{A,i}$) represents the disaggregate (resp. aggregate) number of jobs in queue i.

Disaggregation probabilities

To summarize, for a three queue network, the aggregate transition rates of a given queue i are approximated by Eq. (13). They depend on disaggregation probabilities. We allow for state-dependent probabilities and state-dependent traffic intensities. These probabilities are denoted $\alpha^{\mathrm{f}}_{i,s}$ and $\alpha^{\mathrm{e}}_{i,s}$. We assume them to have the functional form of Eqs. (16) and (17), hence they are given by:

$$\begin{cases} \alpha^{\mathrm{f}}_{i,s} = P(N_i = k_i - 1 \mid N_{A,i} = 1, N_A = s) = \dfrac{(1-\rho_{i,s})}{1-(\rho_{i,s})^{k_i-1}}, & \text{(a)} \\ \alpha^{\mathrm{e}}_{i,s} = P(N_i = 1 \mid N_{A,i} = 1, N_A = s) = \dfrac{(1-\rho_{i,s})(\rho_{i,s})^{k_i-2}}{1-(\rho_{i,s})^{k_i-1}}. & \text{(b)} \end{cases} \tag{22}$$

where N_A is the state vector $(N_{A,i}, N_{A,i+1}, N_{A,i+2})$, and the superscripts e and f refer, respectively, to empty and full (since these expressions are used to approximate the transition rates towards empty and full states, respectively).

Inserting the expression of $\rho_{i,s}$ (Eq. (18)), we obtain:

$$\begin{cases} \alpha^{\mathrm{f}}_{i,s} = \dfrac{(1-\lambda_i/\mu_{i,s})}{1-(\lambda_i/\mu_{i,s})^{k_i-1}}, & \text{(a)} \\ \alpha^{\mathrm{e}}_{i,s} = \dfrac{(1-\lambda_i/\mu_{i,s})(\lambda_i/\mu_{i,s})^{k_i-2}}{1-(\lambda_i/\mu_{i,s})^{k_i-1}}, & \text{(b)} \end{cases} \tag{23}$$

where $\mu_{i,s}$ and λ_i are given by (19) and (20), respectively.

2.3.2. Blocking probabilities

Recall that we describe the state of a queue as either empty, full, or 'non-empty and non-full'. Given a job occupying a server, this state description does not distinguish between a job undergoing service or one that has completed service and is blocked. This section presents a simple approximation of the probability of a job being blocked, i.e., the blocking probability.

The following example allows us to introduce the notion of blocking probability and the complex between-queue dependencies that arise due to blocking. Consider a state $s = (1, 2, 2)$ where queue i (i.e., the most upstream queue) is in (aggregate) state 1, and queues $i+1$ and $i+2$ are in aggregate state 2, i.e. they are full. Assume there is a service completion at queue $i+2$. This service completion can trigger a transition to one of the following states:

- if queue $i+2$ is not blocking queue $i+1$, then the new state is (1, 2, 1);
- if queue $i+2$ is blocking queue $i+1$ and is not blocking queue i, then the new state is (1, 1, 2);
- if queue $i+2$ is blocking queue $i+1$ and is blocking queue i, then the new state can be either (1, 2, 2) or (0, 2, 2).

Table 2
Blocking probabilities.

Blocked queues	Source queue	Initial joint states	Blocking probability
i	$i+1$	({1, 2}, 2, {0, 1})	$\beta_{i,1}$
$i, i+1$	$i+2$	({1, 2}, 2, 2)	$\beta_{i,2}$
$i+1$	$i+2$	({0, 1, 2}, 1, 2), (0, 2, 2)	$\beta_{i,3}$
$i+1$	$i+2$	({1, 2}, 2, 2)	$\beta_{i,4}$

In order to determine the new state to which a transition can take place, we use state-dependent, yet simple and exogenous, approximations for the blocking probabilities. The approximation assumes that service completions follow an exponential distribution. We use the following property of exponential random variables. For a set of n independent service durations $\{X_\ell\}_{\ell=1:n}$, which are exponentially distributed random variables with rate parameter μ_ℓ, the probability that the first service completion is of type ℓ is given by (e.g., Larson and Odoni, 1981, Chapter 2.12.4, Equation (2.62)):

$$P(X_\ell < X_i \quad \forall i \neq \ell) = \frac{\mu_\ell}{\sum_{j=1}^{n} \mu_j}. \tag{24}$$

We use this property to approximate the blocking probabilities. The states that are affected by blocking are listed in Table 2. This table lists the queues that are blocked (column 1), the queue that is at the source of (i.e. causes) the blocking (column 2), the feasible joint states where such blocking can occur (column 3), and the corresponding probability with which this blocking occurs (column 4). For brevity, multiple states for the initial joint states are listed in braces. For instance row 2 accounts for two possible initial states (1,2,2) and (2,2,2). Row 1 considers cases where queue $i+1$ is the source of blocking: queue i (i.e., the most upstream queue of the three queues) is blocked by queue $i+1$. Rows 2 and 4 consider cases where it is queue $i+2$ that is both the source of blocking and could potentially block both upstream queues (because queue $i+1$ is also full). Row 2 considers cases where queue $i+2$ blocks both queues i and $i+1$, while row 4 considers cases where only queue $i+1$ is blocked. Row 3 considers cases where queue $i+2$ is the source of blocking but can only block queue $i+1$ (because queue $i+1$ is not full, so queue i cannot be blocked).

The approximations of the blocking probabilities of Table 2 are given by:

$$\beta_{i,1} = \frac{\mu_i}{\mu_i + \mu_{i+1}} \tag{25}$$

$$\beta_{i,2} = \frac{\mu_i}{\mu_i + \mu_{i+1} + \mu_{i+2}} \frac{\mu_{i+1}}{\mu_{i+1} + \mu_{i+2}} + \frac{\mu_{i+1}}{\mu_i + \mu_{i+1} + \mu_{i+2}} \frac{\mu_i}{\mu_i + \mu_{i+2}} \tag{26}$$

$$\beta_{i,3} = \frac{\mu_{i+1}}{\mu_{i+1} + \mu_{i+2}} \tag{27}$$

$$\beta_{i,4} = \frac{\mu_{i+1}}{\mu_i + \mu_{i+1} + \mu_{i+2}} \frac{\mu_{i+2}}{\mu_i + \mu_{i+2}}. \tag{28}$$

Eqs. (25) and (27) (rows 1 and 3 of Table 2) are derived directly from Eq. (24) and represent the probability that the first (resp. second) queue finishes service before the second (resp. third) queue. That is, blocking occurs due to the queue immediately downstream. Eq. (26) (row 2 of Table 2) considers the scenario where both the first and the second queue are blocked by the third queue. This occurs when both the first and second queues complete their service before the third queue. The equation sums the probabilities of the independent events in which either the first or the second queue finish first out of the three and then the remaining queue of the first two finishes service before the third. Eq. (28) (row 4 of Table 2) considers the case where the second queue is blocked by the third queue, but the first queue is not blocked. This is the probability that the second queue finishes service before the first and third queues and that the next queue to finish service is the third queue.

Let us summarize the procedure to derive the aggregate joint distribution of a three-queue tandem network. The joint distribution $P(N_A)$ is obtained by solving the corresponding (aggregate) global balance Eq. (5), which depend on the transition rate matrix, $\bar{Q}$. The latter is defined as:

$$\bar{Q} = f(\gamma, \mu, k, \alpha^{\mathrm{f}}, \alpha^{\mathrm{e}}, \beta), \tag{29}$$

where γ, μ, and k are exogenous parameters, α^{f} and α^{e} are the state-dependent disaggregation probabilities (defined by Eqs. (19), (20), (21) and (23)), and β denotes the blocking probabilities (Eqs. (25)–(28) and Table 2). The function f maps the parameters of the three-queue network $(\gamma, \mu, k, \alpha^{\mathrm{f}}, \alpha^{\mathrm{e}}, \beta)$ to the transition rate matrix, $\bar{Q}$. This function is tabulated in Table 10 of Appendix A.

For a three queue system, this results in $27 + 12 = 39$ variables, representing the 27 joint states of the system as well as the 12 state-dependent disaggregation probabilities (following (19) there are 3*2 disaggregation probabilities for the first queue, 2*2 for the second queue and 1*2 for the third queue). If we use the proposed model to evaluate a network with 3 links in tandem, then the model corresponds to a squared system of equations with 39 variables and 39 unknowns. For all experiments in this paper, the numerical solver converged to a solution.

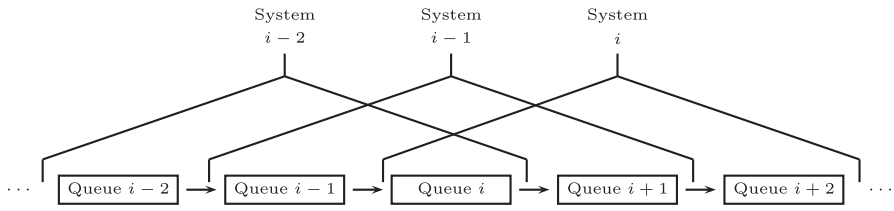

Fig. 3. Overlapping systems of three tandem queues.

2.4. Aggregate description of a tandem network with I queues

This section generalizes the above approach to a tandem network with I queues. If the above approach is directly applied to an I-queue network, the state space would be of dimension 3^I, and would thus increase exponentially with the number of queues. Instead, we view the network as a set of overlapping systems of three tandem queues. The set of I queues is decomposed into $I-2$ overlapping systems each with three queues in tandem. This decomposition is illustrated in Fig. 3. Thus, the number of states is now linear in the number of queues.

Each of the $I-2$ systems can be viewed as a single three-queue system. For each system, the approach of Section 2.3 is applied, i.e. the aggregate joint distribution of each system satisfies the system of equations described in Section 2.3.

Consider a three-queue system within a larger network (e.g., system $i-1$ of Fig. 3). In order to account for its dependencies with adjacent queues, we need to approximate the arrival rate to the most upstream queue, and the effective service rate of its most downstream queue. In the three-queue network of Section 2.3, these two rates were exogenous, however, in a larger tandem network, these rates are now endogenous.

For a given system with queues indexed by $(i, i+1, i+2)$, the arrival rate to the most upstream queue (queue i) is given by the flow conservation Eq. (20). The service rate of the most downstream queue (queue $i+2$) is obtained by following the ideas in Osorio and Bierlaire (2009), where the effective service rate of a queue (which accounts for service and for potential blocking from downstream queues) is given by:

$$\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + p_i^b \frac{1}{\tilde{\mu}_i}, \tag{30}$$

where $\hat{\mu}_i$ denotes the effective service rate, μ_i represents the exogenous service rate, p_i^b denotes the blocking probability, and $\tilde{\mu}_i$ is the unblocking rate. This equation states that the expected effective service time (represented by $1/\hat{\mu}_i$) is the sum of the expected service time (represented by $1/\mu_i$) and the expected blocked time. The latter time is approximated by the probability of a job getting blocked (represented by p_i^b) and the expected blocked time conditional on a job being blocked (represented by $1/\tilde{\mu}_i$). In other words, given that a job is blocked, its expected blocked time is represented by $1/\tilde{\mu}_i$. The term $\tilde{\mu}_i$ is known as the unblocking rate. A further discussion of blocking probabilities and unblocking rates is given, for instance, in Schweitzer and Altiok (1989).

The approximation for the unblocking rate is derived in Osorio and Bierlaire (2009, Eq. (7)). Its simplified expression for a single server queue is detailed in Osorio (2010, Chap. 4) (it is presented with similar notaton to that of this paper in Eq. (3 b) of Osorio and Chong (2015)) and is given by:

$$\frac{1}{\tilde{\mu}_i} = \frac{\lambda_{i+1} P(N_{i+1} < k_{i+1})}{\lambda_i P(N_i < k_i)} \frac{1}{\hat{\mu}_{i+1}}, \tag{31}$$

which is equivalent to:

$$\frac{1}{\tilde{\mu}_i} = \frac{\lambda_{i+1} P(N_{A,i+1} < 2)}{\lambda_i P(N_{A,i} < 2)} \frac{1}{\hat{\mu}_{i+1}}. \tag{32}$$

Eq. (31) (or equivalently Eq. (32)) states that the expected blocked time of queue i is given by the product of the two following terms.

- The expected effective service time of its downstream queue, which is represented by $1/\hat{\mu}_{i+1}$. This states that the expected time between successive unblockings (i.e., between slot availabilities downstream) equals the expected effective service time of the downstream queue. In other words, the unblocking rate at the upstream queue equals the rate at which vehicles can leave the downstream queue.
- The second term can be interpreted, for a tandem network, as the probability of vehicle that arrives from queue i to queue $i+1$ was previously blocked by queue $i+1$.

The probability that queue i is blocked, p_i^b, is approximated by:

$$p_i^b = P(N_{A,i+1} = 2) \frac{\mu_i}{\mu_i + \mu_{i+1}}, \tag{33}$$

which considers the probability that the immediately downstream queue is full and the probability that the upstream queue finishes service before the downstream queue. This approximation follows the same reasoning as Eq. (24).

The model of this paper is formulated for tandem topology networks, where each queue has a single downstream queue. An extension for general topology networks can be obtained as in Osorio (2010, Chap. 4) and in Osorio and Bierlaire (2009). More specifically, for general topology networks, traffic on a queue may be blocked due to a spillback from a downstream queue occupying the physical space of the intersection, and hence blocking traffic from all upstream links regardless of their downstream destination links. This can be accounted for by using an extension of Eq. (33) (e.g., one can account for the blocking probabilities of all downstream links).

To summarize, Eqs. (30), (32) and (33) define the approximation of the effective service rates ($\hat{\mu}$) as a function of the aggregate probabilities ($P(N_A)$), the arrival rates (λ) and the exogenous service rates (μ).

Thus, when analyzing a three-queue system (with queues indexed by $(i, i+1, i+2)$) that is embedded within a larger network of queues, we apply the procedure of Section 2.3, with the arrival to queue i given by Eq. (20) and the effective service rate of queue $i+2$, $\hat{\mu}_{i+2}$, given by Eqs. (30), (32) and (33).

Since the systems overlap (see Fig. 3), there are individual and pairs of queues that are common to multiple systems. We include the following system of linear constraints in order to ensure that the one or two-dimensional marginal distributions obtained are equivalent regardless of the system from which they are obtained.

For an I-queue network, there are $I-3$ pairs of queues that are contained in two different systems. This leads to 3^2 equations (a pair of queues has 3^2 joint states) for each overlapping pair of queues.

$$\sum_{a=0}^{2} P_{i-1}(N_{A,i-1}=a, N_{A,i}=b, N_{A,i+1}=c) = \sum_{d=0}^{2} P_i(N_{A,i}=b, N_{A,i+1}=c, N_{A,i+2}=d),\ (b,c)\in\{0,1,2\}^2,\ i\in\{2,\ldots,I-2\} \tag{34}$$

where P_i denotes the distribution obtained from analyzing the i^{th} three-queue system. Enforcing consistency between systems with 2 overlapping queues also ensures consistency between systems with 1 overlapping queue.

Implementation notes

For a network of I queues, there are $I-2$ three-queue systems. We implement the model with $41(I-2)+I$ variables, which consist of $27(I-2)$ joint state probabilities, $12(I-2)$ disaggregation probabilities ($\alpha^{\text{e}}, \alpha^{\text{f}}$), $I-2$ arrival rates of the most upstream queue of each system (λ), $I-2$ effective service rates of the most downstream queue of each system ($\hat{\mu}$) and I probabilities of a queue being full (one for each queue) ($P(N_{A,i}=2)$). Additionally, there are $9(I-3)$ equations that ensure consistency. Thus, the model consists of a total of $41(I-2)+I+9(I-3)$ equations.

3. Validation

We compare the aggregate joint queue-length distributions obtained by the proposed model with those estimated from a discrete event simulation model of a Markovian FCQN (Meier, 2007). For all simulation experiments in this section, 10,000 simulation replications are run, each with a duration of 1000 time units in order to ensure stationarity. The time unit is defined by the temporal unit of the arrival and service rates. It can be chosen arbitrarily. For each replication, the disaggregate state is evaluated at time 1,000, from which we derive the aggregate state.

Note that for an isolated M/M/1/k queue, the transient distribution converges to the stationary distribution with an exponential decay of the form e^{-xt}, where $x\in\{\lambda+\mu\pm 2\sqrt{\lambda\mu}\}$ (Morse, 1958, pages 65–67), which for these scenarios is in the order of 10 time units. Hence, stationarity is guaranteed within 1000 time units. Note also that for finite capacity networks, stationarity is guaranteed for all levels of traffic intensity.

Let p_s denote the stationary probability of being in a given aggregate state s. A 95% confidence interval for p_s is given by: $\hat{p}_s \pm 1.96\sqrt{\frac{\hat{p}_s(1-\hat{p}_s)}{10{,}000-1}}$ (see, for instance, Section 7.3.3 of Rice (1994)), where $\hat{p}_s$ is the simulated estimate of p_s. This confidence interval is displayed as error bars in the figures of this section.

The analytical approximation of the aggregate joint distribution is derived by using the nonlinear system of equations solver of Matlab ("Levenberg-Marquardt" algorithm of the *fsolve* solver) (Mathworks, Inc., 2012) with a tolerance of 10^{-6}. The initial joint distribution (i.e., initial point) provided to the solver is the product of approximate marginal distributions, which are each obtained by solving the FCQN model given in Appendix C.

3.1. Three-queue network

We assume that external arrivals only occur at the first (most upstream) queue, with $\gamma_1 = 1.8$. We consider 27 scenarios with differing service rates and queue capacities, displayed in Table 3. We refer to the queue with the highest traffic intensity as the bottleneck queue. The 27 experiments are defined by varying: (i) the queue capacities (which can take values: {2,5,10}), (ii) the bottleneck queue traffic intensity (which can take values {0.9, 0.6, 0.45}), (iii) the location of the bottleneck (which can be either at the most upstream queue, the most downstream queue, or all 3 queues).

In scenarios 1–9, the minimum service rate is 2, implying a bottleneck queue traffic intensity of 0.9. Scenarios 10–18 (resp. 19–27), are identical to scenarios 1–9 except for the value of the minimum service rate which is 3 (resp. 4) instead

Table 3
Three-queue network scenarios.

Scenario	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
μ_1	2	2	2	2	2	2	6	6	6	3	3	3	3	3	3	6	6	6	4	4	4	4	4	4	6	6	6
μ_2	2	2	2	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
μ_3	2	2	2	6	6	6	2	2	2	3	3	3	6	6	6	3	3	3	4	4	4	6	6	6	4	4	4
k_1	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_2	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_3	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
Set (Fig. 6)	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
Set (Fig. 7)	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	9
Set (Fig. 8)	1	2	3	1	2	3	1	2	3	4	5	6	4	5	6	4	5	6	7	8	9	7	8	9	7	8	9

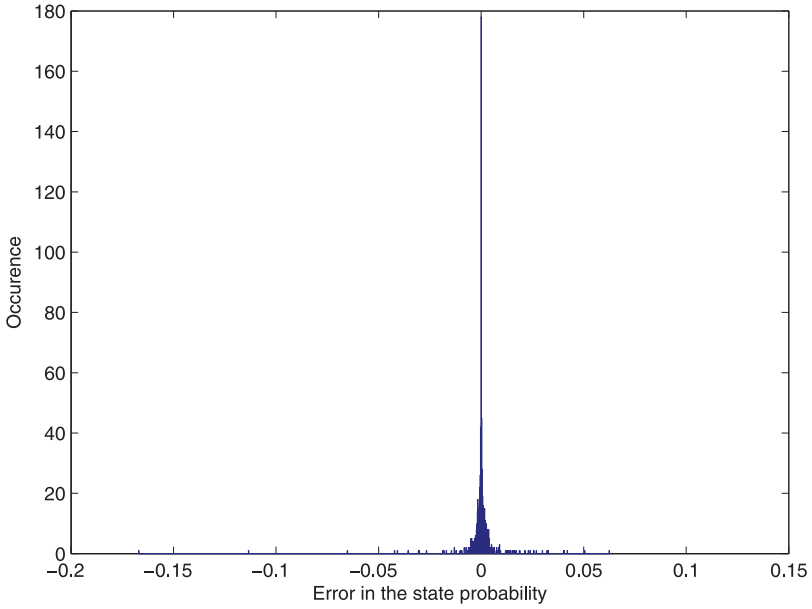

Fig. 4. Histogram of the approximation errors for the three-queue network experiments.

of 2, this leads to a bottleneck queue traffic intensity of 0.6 (resp. 0.45). In the figures hereafter the different scenarios are grouped into sets. The last 3 rows of Table 3 indicate, for each figure, the set that each scenario belongs to.

Fig. 4 displays the histogram of the errors for all joint state probabilities, this considers all joint states and all scenarios. The error is calculated as the difference between the simulated estimate and the analytical estimate. The average of the 729 absolute errors is 0.0032, with a standard deviation of 0.0099. Fig. 5 displays the 729 state probability values. The x-axis (resp. y-axis) considers the simulated (resp. analytical) estimates. The diagonal line $x = y$ is also displayed. Fig. 5 indicates that the vast majority of the points lie along this line. This figure gives, for each state probability, both absolute and relative error information. Overall, the proposed method yields very accurate approximations.

In order to compare the performance of the proposed method across scenarios, we calculate, for each scenario, the average absolute error (AAE) between all joint state probabilities. The average is taken over all possible states. In a three-queue network, the joint aggregate distribution has 3^3 states, hence the average is taken over these 27 probabilities.

Fig. 6 displays the value of the AAE for each scenario. The scenarios are grouped into 9 sets (the x-axis gives the set index). The set that each scenario belongs to is indicated in the third to last line of Table 3. A given scenario set contains 3 scenarios, which are identical except for the value of the traffic intensity of the bottleneck queue. In other words, for a given x value, the 3 points correspond to scenarios with common space capacity values and common bottleneck location, but different traffic intensity for the bottleneck queue.

In this figure, the scenarios indexed 1–9 (resp. 10–18, and 19–27) of Table 3 correspond to the circles (resp. crosses and squares) indexed by the sets 1–9. This figure shows that for 7 out of the 9 scenario sets (all sets but 4 and 5), the AAE increases with the bottleneck traffic intensity.

Sets 1–3 consider the scenarios where all 3 queues are bottleneck queues. Set 1 (resp. 2 and 3) considers space capacities of 2 (resp. 5 and 10). For these sets, the AAE for the scenarios with bottleneck traffic intensities of 0.45 and 0.6 are similar. For bottleneck traffic intensities of 0.9, the AAE increases with the space capacity. Sets 7–9 consider the scenarios where the bottleneck queue is the most downstream queue. Set 7 (resp. 8 and 9) considers space capacities of 2 (resp. 5 and 10). The

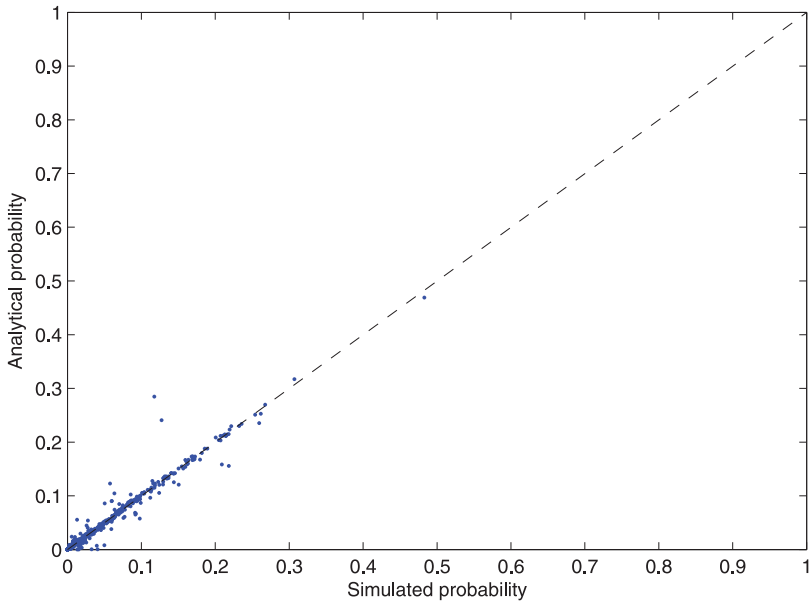

Fig. 5. State probabilities for the three-queue network experiments: simulated estimates versus analytical approximations.

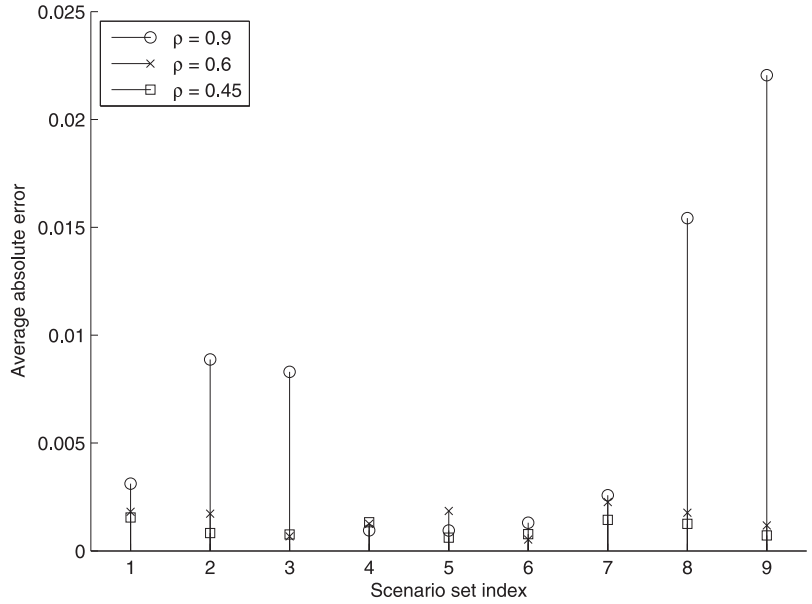

Fig. 6. Three-queue network: impact of the bottleneck queue traffic intensity on the average absolute error.

same conclusions as for sets 1–3 hold: the AAE for the scenarios with bottleneck traffic intensities of 0.45 and 0.6 are similar, while for bottleneck traffic intensities of 0.9, the AAE increases with the space capacity. As the space capacity increases, so does the number of states that are aggregated into the aggregate state ($N_A = 1$). Hence, it becomes more challenging to approximate the disaggregation probabilities.

Sets 4–6 present a different trend. They consider the scenarios where the bottleneck queue is the most upstream queue. Set 4 (resp. 5 and 6) considers space capacities of 2 (resp. 5 and 10). For these sets, the AAE for all 3 bottleneck traffic intensity values (0.9, 0.6, 0.45) are similar, and they do not vary much with the space capacity values. When the most upstream queue is the bottleneck queue, blocking has a low probability of occurrence. In other words there are no complex within-network congestion effects. Hence, the between-queue interactions are not as intricate to approximate analytically.

Fig. 7 also displays the value of the AAE for each of 9 scenario sets. A given scenario set contains the 3 scenarios, which are identical except for the value of the space capacity. The set that each scenario belongs to is indicated in the second to last line of Table 3. The sets 4–9 have very similar AAE values, and their AAE values do not vary much with space

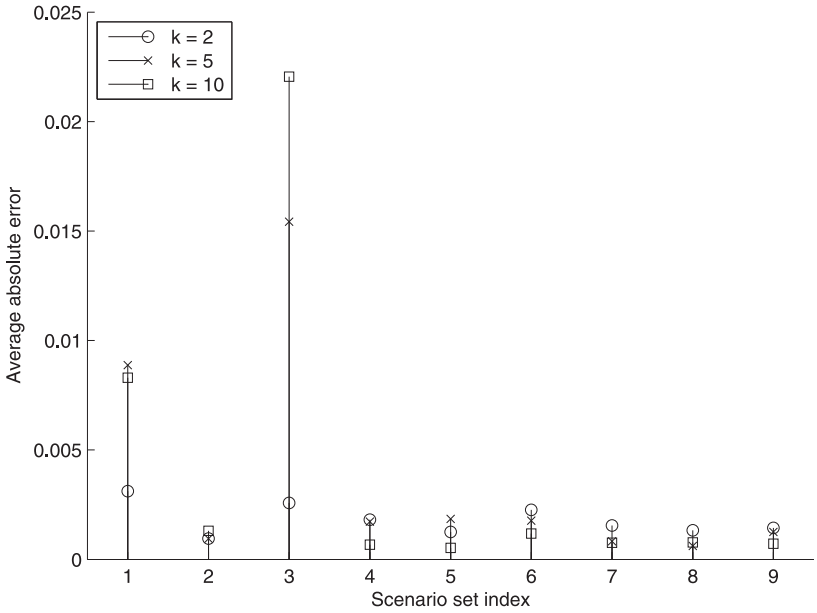

Fig. 7. Three-queue network: impact of the space capacity on the average absolute error.

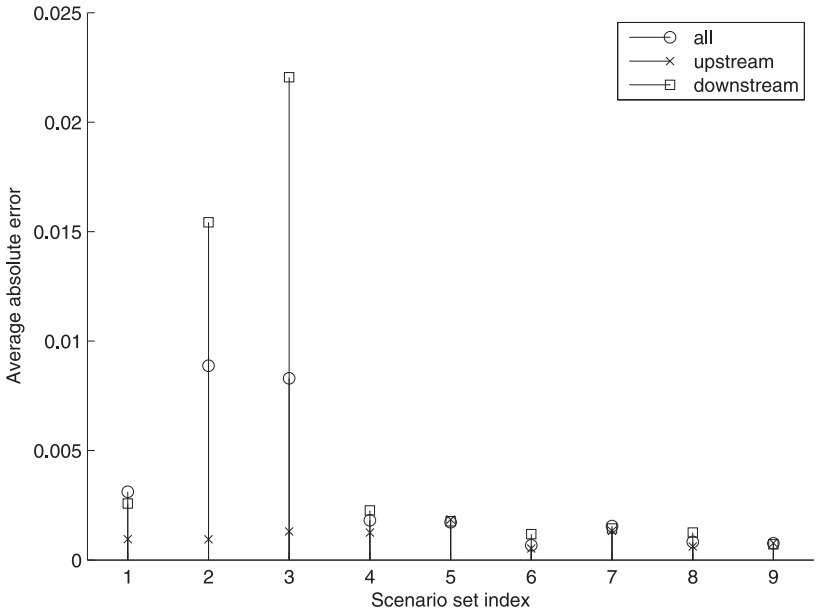

Fig. 8. Three-queue network: impact of the location of the bottleneck on the average absolute error.

capacity. These sets correspond to scenarios 10–27, where the bottleneck traffic intensity is not very high. In other words, for scenarios without heavily congested traffic the accuracy of the proposed method does not vary with the space capacity values. The sets 1–3 group the scenarios 1–9, where congestion is high. Set 2 contains the scenarios where the bottleneck is located at the most upstream queue. As was also indicated in the analysis of Fig. 6, for these scenarios blocking has a low probability of occurrence, and hence the between-queue interactions are less intricate to approximate analytically. For sets 1 and 3 the AAE varies with space capacity, and tends to increase with the space capacity. As mentioned for Fig. 6, as the space capacity increases so does the number of states aggregated, this makes the approximation of the disaggregation probability more difficult. Set 3 considers the scenarios where the bottleneck queue is the most downstream queue. These are the scenarios, where blocking is the most likely. Hence, for these scenarios the between-queue interactions are the most difficult to approximate.

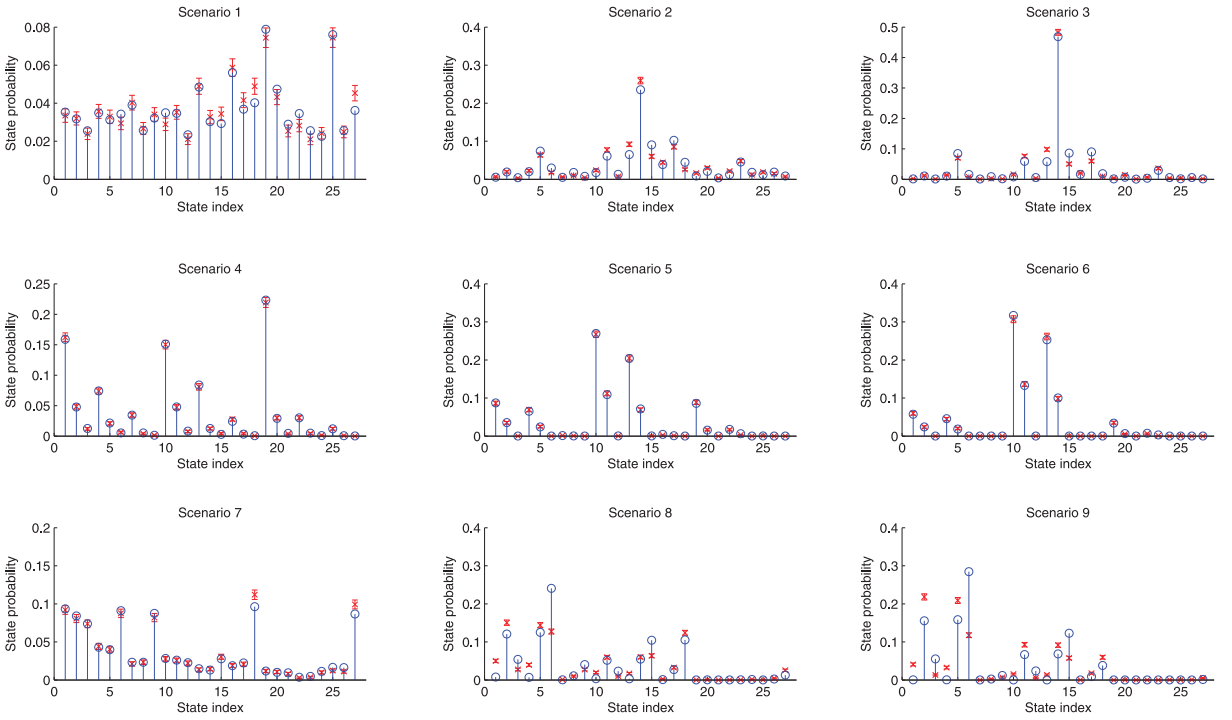

Fig. 9. Aggregate joint distribution for scenarios 1–9 of the three-queue network. The blue circles represent the proposed model approximations, and the red crosses represent the simulation estimates with their corresponding error bars.

Fig. 8 displays the value of the AAE for each of 9 scenario sets. A given scenario set contains the 3 scenarios, which are identical except for the location of the bottleneck queues. The set that each scenario belongs to is indicated in the last line of Table 3. As for Fig. 7, the sets 4–9 have very similar AAE values, and their AAE values do not vary much with the bottleneck location. These sets correspond to scenarios 10–27, where the bottleneck traffic intensity is not very high. As for Fig. 7, for scenarios without heavily congested traffic the accuracy of the proposed method does not vary with the bottleneck location. Sets 1–3 consist of all the scenarios with highly congested traffic. If the bottleneck location is the most upstream queue (displayed as crosses in the figure), the AAE does not vary. This is because congestion does not propagate further than the most upstream queue, hence there is a low probability of blocking, which leads to less intricate to approximate between-queue interactions.

Set 1 (resp. 2 and 3) contains scenarios {1, 4, 7} (resp. {2, 5, 8} and {3, 6, 9}) where the space capacity is 2 (resp. 5 and 10). All scenarios of set 1 have a space capacity of 2, where the aggregate state consists of a singleton state. Hence, there is no need to approximate the disaggregation probabilities. This leads to more accurate approximations, than for sets 2 and 3.

Sets 2 and 3, consider the scenarios with high congestion and where both the disaggregation probabilities and the blocking probabilities need to be approximated. For these scenarios, the location of the bottleneck impacts the AAE. The AAE increases as the blocking probabilities increase. In other words the AAE is lowest for an upstream bottleneck (where blocking occurs with very low probabilities), followed by the scenarios where all queues are bottlenecks, and finally where the most downstream queue is the only bottleneck. This indicates that for networks with high congestion: the analytical approximations are more accurate if the traffic intensities of the queues are similar, rather than if there are isolated locations with higher traffic intensities.

We consider the 9 scenarios with the highest congestion, and hence high AAE values, these are scenarios indexed 1–9 in Table 3. Fig. 9 considers one plot per scenario. Each plot displays the aggregate joint distribution for the scenario. The blue circles represent the proposed approximations, and the red crosses represent the simulation estimates with their corresponding error bars. This figure shows that for most scenarios the analytical approximation of the joint distribution is highly accurate (the analytical estimates are within the confidence intervals).

Note that the AAE value for scenario i ($i \in \{0, 1, \ldots, 9\}$) is displayed in Fig. 6 as the circle with x-axis value of i. From Fig. 6, we can see that the scenarios with highest AAE values correspond to scenarios 2, 3, 8 and 9. Fig. 9 shows that for these scenarios, the form of the joint aggregate distribution is captured by the analytical approximation.

In scenarios 7 - 9, blocking is most likely to occur as a result of the most downstream queue being the bottleneck queue. This leads to the most complex blocking configurations. For these three scenarios, the accuracy of the proposed approximation decreases as the space capacity increases. This can be partly explained by the increasing difficulty to approximate the disaggregation probabilities as the space capacity increases. For scenario 7, the space capacity equals 2, hence the aggregate

Table 4
Six-queue network scenarios.

Scenario	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
μ_1	2	2	2	2	2	2	6	6	6	3	3	3	3	3	3	6	6	6	4	4	4	4	4	4	6	6	6
μ_2	2	2	2	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
μ_3	2	2	2	6	6	6	2	2	2	3	3	3	6	6	6	3	3	3	4	4	4	6	6	6	4	4	4
μ_4	2	2	2	2	2	2	6	6	6	3	3	3	3	3	3	6	6	6	4	4	4	4	4	4	6	6	6
μ_5	2	2	2	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
μ_6	2	2	2	6	6	6	2	2	2	3	3	3	6	6	6	3	3	3	4	4	4	6	6	6	4	4	4
k_1	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_2	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_3	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_4	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_5	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_6	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
Set (Fig. 12)	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
Set (Fig. 13)	1	2	3	1	2	3	1	2	3	4	5	6	4	5	6	4	5	6	7	8	9	7	8	9	7	8	9
Set (Fig. 14)	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	9

state is a singleton state, and the approximation of the disaggregation probability is exact. Scenarios 8 and 9, consider larger space capacity values and cases where a congested queue has an upstream queue with a much smaller traffic intensity. Hence, they illustrate the complexity of accurately approximating both the intricate blocking effects between queues, as well as their interaction with the within-queue disaggregate states (i.e., the disaggregation probabilities). Thus, the following validation scenarios will consider these complex settings where both the space capacity is large and significant blocking occurs within the network.

3.2. Six-queue network

We consider an extension of the three-queue network experiments. The scenarios of the six-queue network are the same as for the three-queue network, but with 2 sets of three-queue networks in tandem. The detailed configurations of the 27 scenarios are tabulated in Table 4. In other words, for scenario i of the six-queue network queue 1,2,3,4,5, and 6, respectively, has the same configuration as queue 1,2,3,1,2 and 3 of scenario i of the three-queue network.

The 27 experiments are defined just as for the three-queue network in that they vary: (i) the queue capacities (which take values: {2,5,10}), (ii) the bottleneck queue traffic intensity (which takes values {0.9, 0.6, 0.45}). Recall that the results of the three-queue network showed that scenarios where a congested queue has an upstream queue with a much lower traffic intensity lead to a high probability of blocking, and hence an increased difficulty in analytically describing the between-queue interactions. In the scenarios of the six-queue network the bottleneck can be located either (i) at queues {1, 4}, (ii) at queues {3, 6} or (iii) at all queues. Hence, these scenarios consider adjacent queues with very different traffic intensities. In other words, the considered bottleneck locations of these scenarios lead to the most complex between-queue interactions. As for the three-queue network, we present an analysis that groups the different scenarios into sets. The last 3 rows of Table 4 indicate, for each figure, the set that each scenario belongs to.

Fig. 10 displays the histogram of the errors for all joint state probabilities, this considers all joint states and all scenarios. The average of the 2916 absolute errors is 0.0079, with a standard deviation of 0.022. Fig. 11 displays the 2916 state probability values. The simulated estimates are along the x-axis, while the analytical approximations are along the y-axis. The diagonal line $x = y$ is displayed. This figure indicates that the vast majority of the points lie along this diagonal line. This figure gives information about both the absolute and the relative errors of each state probability. Overall, the proposed method yields very accurate approximations.

Fig. 12 displays the value of the AAE for each scenario. The scenarios are grouped into 9 sets, which are identical except for the value of the traffic intensity of the bottleneck queue. The set that each scenario belongs to is indicated in the third to last line of Table 4. This figure indicates that as the traffic intensity of the bottleneck queue increases, so does the AAE.

Fig. 13 displays the value of the AAE for each of 9 scenario sets. A given scenario set contains the 3 scenarios, which are identical except for the location of the bottleneck queues. The set that each scenario belongs to is indicated in the second to last line of Table 4. The circles denote the scenarios where all queues are bottleneck queues. The crosses (resp. squares) denote the scenarios where queues 1 and 4 (resp. 3 and 6) are bottleneck queues. Just as for the three-queue network, the smallest AAE values are observed for the cases where the most upstream queue of the network (queue 1) is a bottleneck queue and is followed by a non-bottleneck queue (queue 2). This leads to a congested upstream queue, and consequently to lost opportunities to process jobs, to limited within-network congestion, and to limited occurrence of blocking.

Sets 1–3 (resp. 4–6 and 7–9) contain all scenarios with a bottleneck traffic intensity value of 0.9 (resp. 0.6 and 0.45). This figure shows that as long as the network is not highly congested, the AAE does not vary much with the bottleneck location. For highly congested scenarios (i.e., sets 1–3), the scenarios of set 1 (resp. sets 2 and 3) have space capacities equal to 2 (resp. 5 and 10). For a given bottleneck location the AAE values increase from sets 1 to 2 and to 3. Hence, for highly

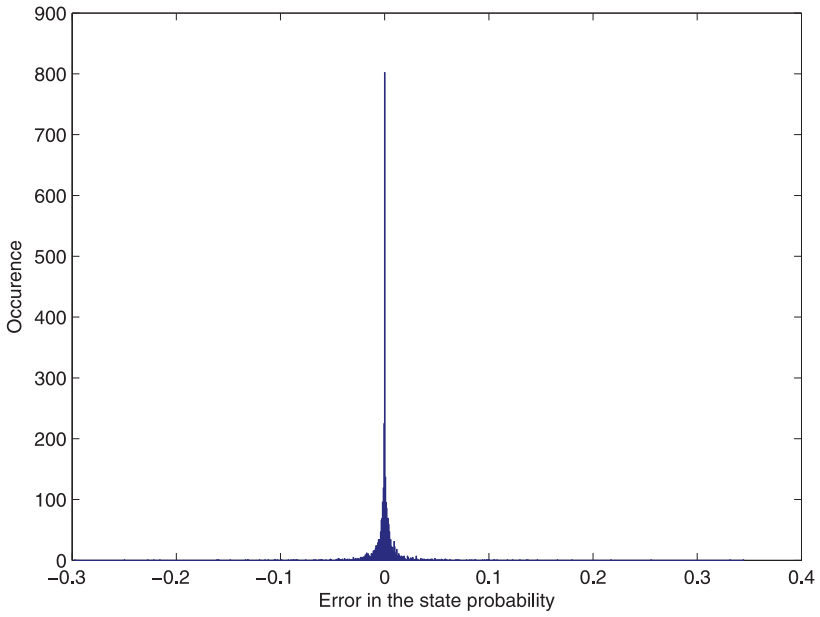

Fig. 10. Histogram of the approximation errors for the six-queue network experiments.

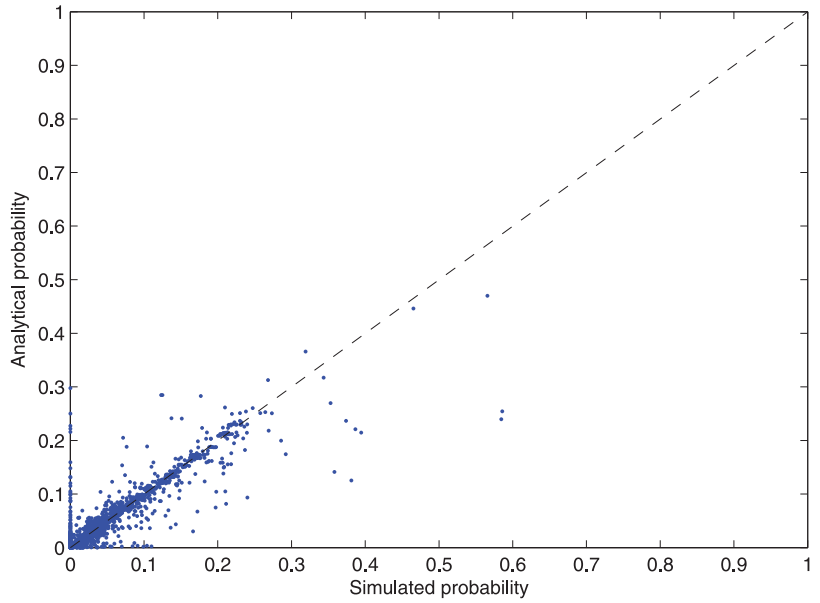

Fig. 11. State probabilities for the six-queue network experiments: simulated estimates versus analytical approximations.

congested scenarios, the AAE increases with increasing space capacity values. This may be due to the added difficulty of approximating the disaggregation probabilities under highly-congested conditions.

Fig. 14 displays the value of the AAE for each of 9 scenario sets. A given scenario set contains the 3 scenarios, which are identical except for the value of the space capacity. The set that each scenario belongs to is indicated in the last line of Table 4. Sets 1–3 (resp. 4–6 and 7–9) contain all scenarios with a bottleneck traffic intensity value of 0.9 (resp. 0.6 and 0.45). The sets with the highest bottleneck traffic intensity value (sets 1–3) lead to the highest AAE values. For these sets, the AAE increases as the space capacity increases. This may be due to an increasing difficulty of approximating the disaggregation probabilities under congested conditions.

For the sets with smaller bottleneck traffic intensity values (sets 4–9), the trend is inversed: the AAE increases as the space capacity decreases. This may be because the blocking probability increases as the space capacity increases. Hence, the scenarios with small space capacities, have higher occurrence of blocking and hence a greater difficulty in analytically

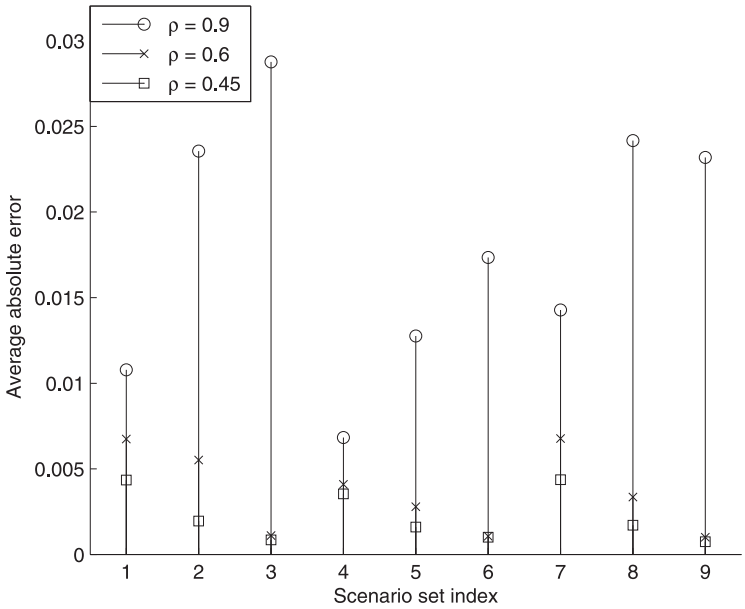

Fig. 12. Six-queue network: impact of the bottleneck queue traffic intensity on the average absolute error.

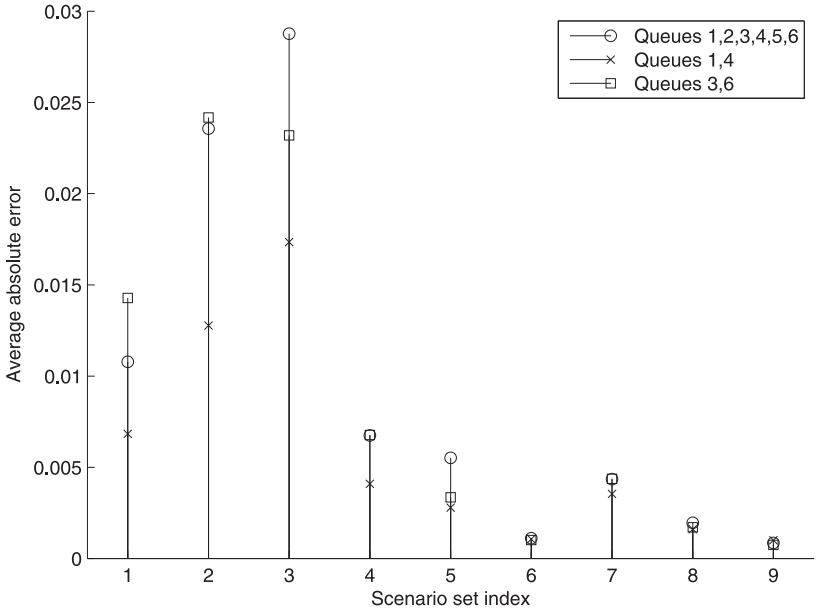

Fig. 13. Six-queue network: impact of the bottleneck location on the average absolute error.

describing the effects of blocking. For sets 4–9 where congestion is not high, for a given space capacity value, the AAE values are similar across sets. In other words, the bottleneck traffic intensities and the bottleneck locations do not impact much the AAE.

3.3. Nine-queue network

We extend the six-queue network into a nine-queue network, in the same way we extended the three-queue network into a six-queue network. In other words, the scenarios of the nine-queue network are the same as for the three-queue network, but with 3 sets of three-queue networks in tandem. The detailed configurations of the 27 scenarios are tabulated in Table 5. In other words, for scenario i of the nine-queue network queues 1–9 have, respectively, the same configuration as queues 1,2,3,1,2,3,1,2 and 3 of scenario i of the three-queue network.

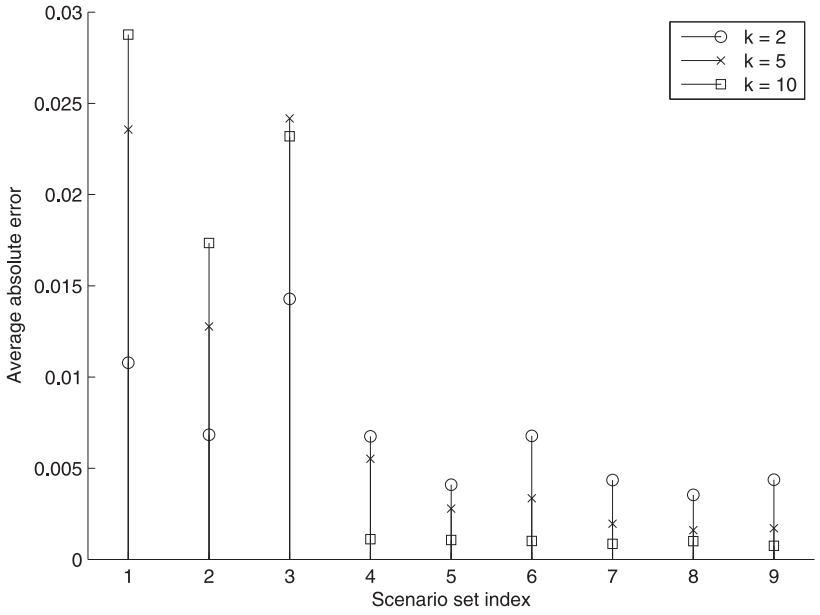

Fig. 14. Six-queue network: impact of the space capacity on the average absolute error.

Table 5
Nine-queue network scenarios.

Scenario	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
μ_1	2	2	2	2	2	2	6	6	6	3	3	3	3	3	3	6	6	6	4	4	4	4	4	4	6	6	6
μ_2	2	2	2	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
μ_3	2	2	2	6	6	6	2	2	2	3	3	3	6	6	6	3	3	3	4	4	4	6	6	6	4	4	4
μ_4	2	2	2	2	2	2	6	6	6	3	3	3	3	3	3	6	6	6	4	4	4	4	4	4	6	6	6
μ_5	2	2	2	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
μ_6	2	2	2	6	6	6	2	2	2	3	3	3	6	6	6	3	3	3	4	4	4	6	6	6	4	4	4
μ_7	2	2	2	2	2	2	6	6	6	3	3	3	3	3	3	6	6	6	4	4	4	4	4	4	6	6	6
μ_8	2	2	2	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
μ_9	2	2	2	6	6	6	2	2	2	3	3	3	6	6	6	3	3	3	4	4	4	6	6	6	4	4	4
k_1	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_2	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_3	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_4	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_5	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_6	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_7	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_8	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
k_9	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10	2	5	10
Set (Fig. 17)	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
Set (Fig. 18)	1	2	3	1	2	3	1	2	3	4	5	6	4	5	6	4	5	6	7	8	9	7	8	9	7	8	9
Set (Fig. 19)	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	9

The 27 experiments are defined just as for the three-queue network: (i) same values for the space capacities, (ii) same values for the bottleneck queue traffic intensities. In the scenarios of the nine-queue network the bottleneck can be located either (i) at queues {1, 4, 7}, (ii) at queues {3, 6, 9} or (iii) at all queues. Hence, these scenarios consider adjacent queues with very different traffic intensities. In other words, the considered bottleneck locations of these scenarios lead to the most complex between-queue interactions. The last 3 rows of Table 5 indicate, for each figure, the set that each scenario belongs to.

Fig. 15 displays the histogram of the errors for all joint state probabilities, this considers all joint states and all scenarios. The average of the 5103 absolute errors is 0.0086, with a standard deviation of 0.022. Fig. 16 displays the 5103 state probability values. The x-axis (resp. y-axis) considers the simulated (resp. analytical) estimates. The diagonal line $x = y$ is also displayed. Fig. 16 indicates that the vast majority of the points lie along or near this line. Overall, the proposed method yields very accurate approximations.

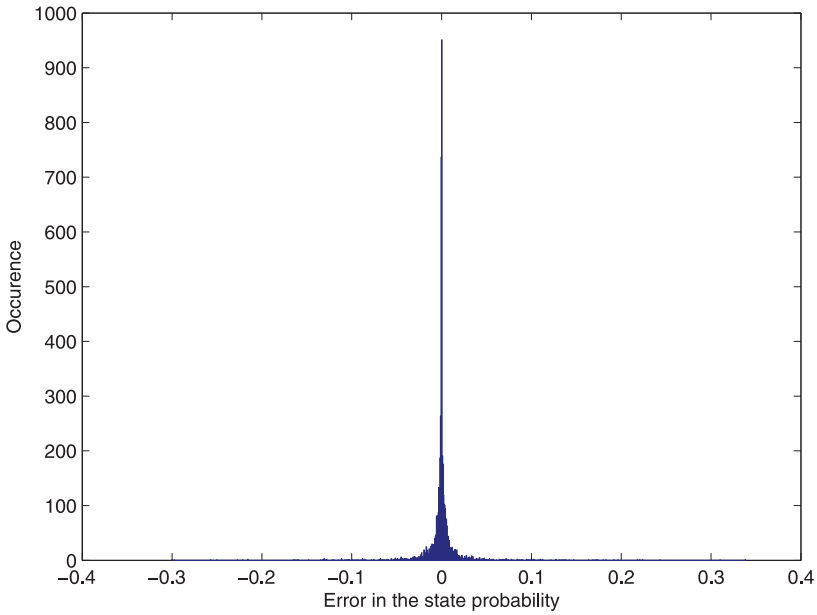

Fig. 15. Histogram of the approximation errors for the nine-queue network experiments.

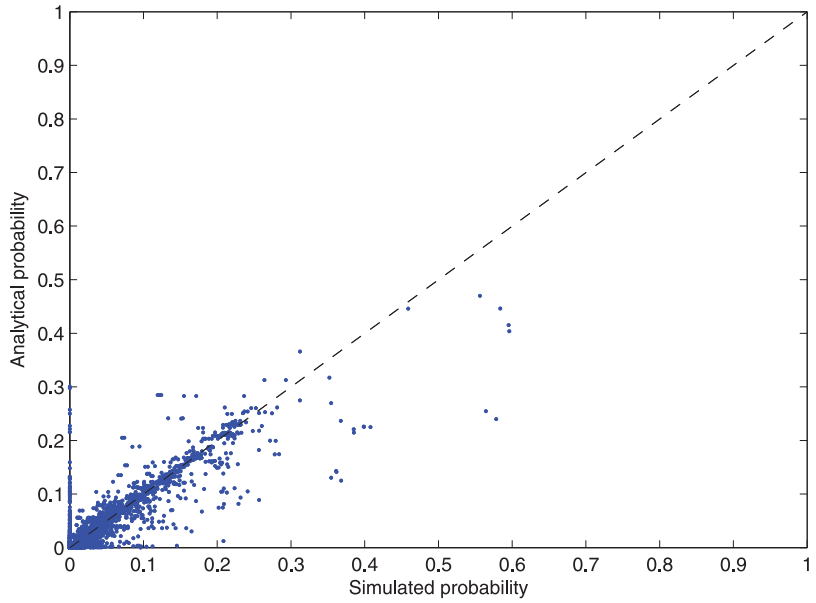

Fig. 16. State probabilities for the nine-queue network experiments: simulated estimates versus analytical approximations.

Fig. 17 displays the value of the AAE for each scenario. The scenarios of a given scenario set are identical except for the value of the traffic intensity of the bottleneck queue. The set that each scenario belongs to is indicated in the third to last line of Table 5. This figure indicates that as the traffic intensity of the bottleneck increases, so does the AAE.

Fig. 18 displays the value of the AAE for each of 9 scenario sets. The scenarios of a given scenario set are identical except for the location of the bottleneck queues. The set that each scenario belongs to is indicated in the second to last line of Table 5. The circles denote the scenarios where all queues are bottleneck queues. The crosses (resp. squares) denote the scenarios where queues 1, 4 and 7 (resp. 3, 6 and 9) are bottleneck queues. This figure exhibits the same trends as for the six-queue network (Fig. 13), the same conclusions hold.

Fig. 19 displays the value of the AAE for each of 9 scenario sets, where a given scenario set contains the scenarios, which are identical except for the value of the space capacity. The set that each scenario belongs to is indicated in the last line of Table 5. Again, this figure exhibits the same trends as for the six-queue network (Fig. 18), the same conclusions hold.

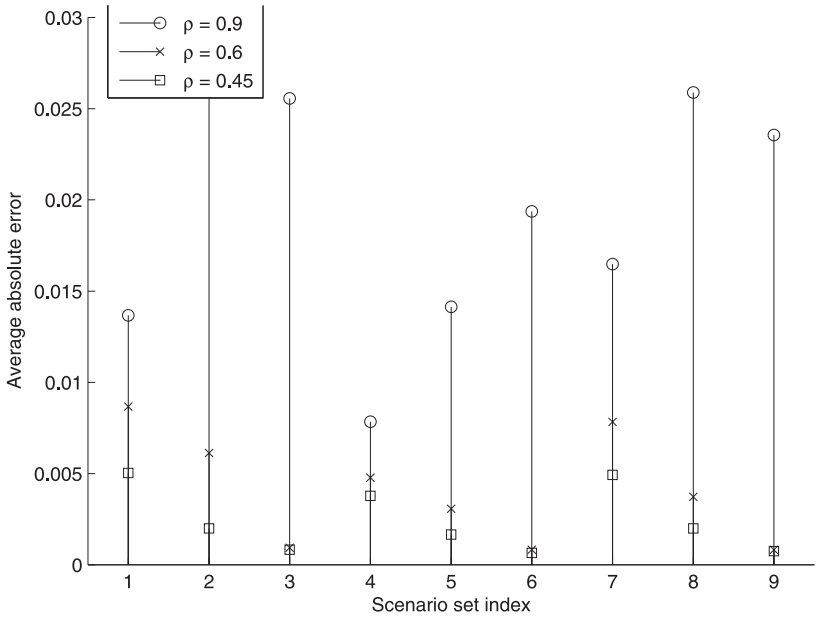

Fig. 17. Nine-queue network: impact of the bottleneck queue traffic intensity on the average absolute error.

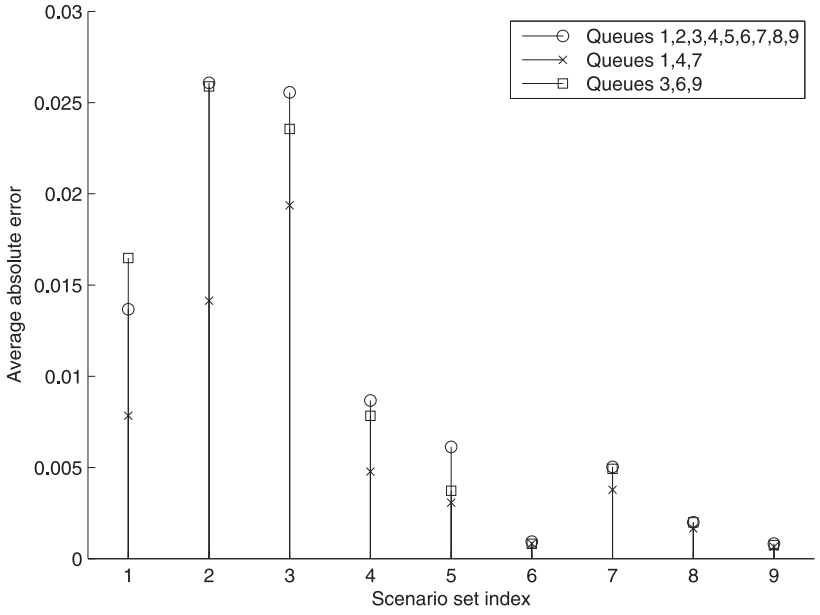

Fig. 18. Nine-queue network: impact of the bottleneck location on the average absolute error.

3.4. Twenty-four queue network

We consider a tandem network with 24 queues. For all queues $\mu_i = 10$ and $\gamma_i = 0.4$. This leads to a traffic intensity of queue i equal to $0.4^{*}i/10$, which ranges from 0.004 to 0.96. This configuration leads to all queues having larger traffic intensity than their corresponding upstream queues. As mentioned, above this increases the occurrence of blocking, and leads to scenarios that are analytically challenging to approximate. All queues have the same space capacity, with 9 space-capacity scenarios defined in Table 6.

Fig. 20 displays the histogram of the errors for all joint state probabilities, this considers all joint states and all scenarios. The average of the 5346 absolute errors is 0.0351, with a standard deviation of 0.099. Overall, the proposed method yields very accurate approximations.

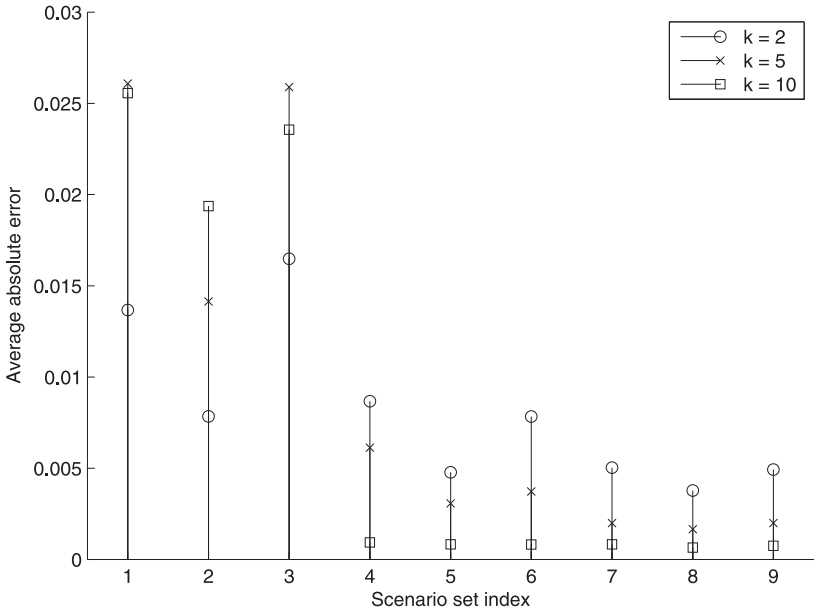

Fig. 19. Nine-queue network: impact of the space capacity on the average absolute error.

Table 6
Space capacity scenarios for the 24-queue network.

Scenario	1	2	3	4	5	6	7	8	9
k_i	5	10	15	20	25	30	35	40	45

Table 7
Space capacity scenarios proposed by Bell (1982).

Scenario	1	2	3	4	5	6	7	8	9
k_1	2	2	3	3	3	4	5	6	11
k_2	2	3	2	3	4	4	5	6	11

Fig. 21 displays the AAE for each scenario. Notice that as the space capacity increases (i.e., as the scenario index increases), the occurrence of blocking decreases (hence it is easier to analytically describe the between-queue dependencies), but the number of states within the aggregate state increases (hence it is more difficult to approximate the disaggregation probabilities). Fig. 21 indicates that the AAE does not vary with the space capacity. Hence, the proposed method leads to accurate approximations of the disaggregation probabilities, even for high-dimensional aggregate states.

3.5. Expected throughput upper bound

A theoretical upper bound on the expected throughput rate of M/M/c/K networks is derived in Bell (1982). The latter work showed that several decomposition methods "lead to impossible mean throughput rates". It considered a two-queue single server network, $\mu_1 = 3, \mu_2 = 1, \gamma_1 = 1$ and $\gamma_2 = 0$. The considered scenarios vary the space capacity of the queues, they are displayed in Table 7. Fig. 22 compares the upper bound of Bell (1982) to the expected throughput of the decomposition methods of Singh and Smith (1997), Kerbache and Smith (1988), Boxma and Konheim (1981), Takahashi et al. (1980), Hillier and Boling (1967) and our proposed method. The x-axis (resp. y-axis) represents the scenario index (resp. the expected network throughput).

In our method, the network throughput is approximated by $\gamma_1 P(N_1 > 0)$. Of these methods, those that yield reasonable throughput approximations (i.e., they do not significantly violate the bounds) are the methods of Singh and Smith (1997), Kerbache and Smith (1988) and our proposed method. Table 8 displays the numerical values of the expected throughput for each of these three methods, as well as the bound values. For scenarios 2, 5–9 the proposed method satisfies the bounds. For scenarios 1, 3 and 4 the bound is violated by 1.6%, 3.7% and 0.2%, respectively. Overall, the proposed method yields reasonable values for the expected network throughput.

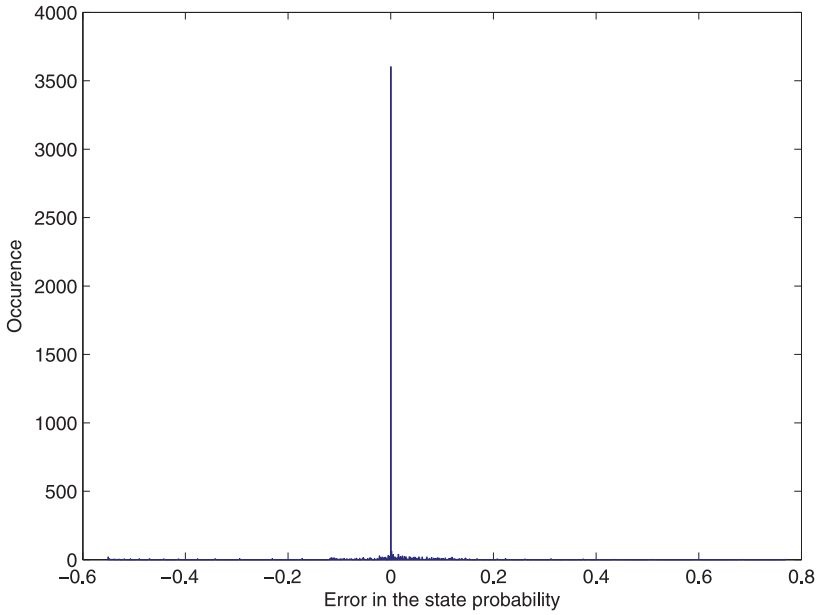

Fig. 20. Histogram of the approximation errors for the 24-queue network experiments.

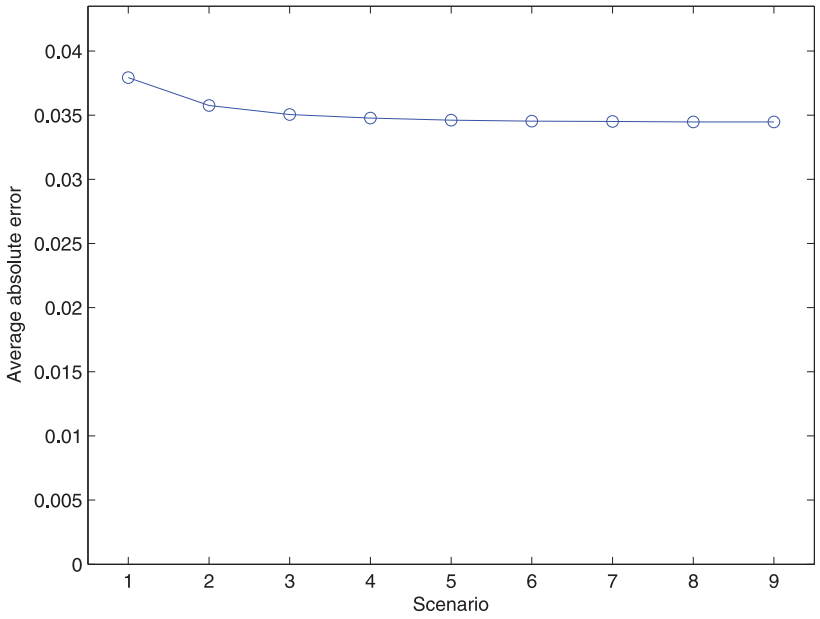

Fig. 21. Impact of increasing space capacities on the average absolute error the 24-queue network.

Table 8
Expected throughput approximations and upper bound.

Scenario	1	2	3	4	5	6	7	8	9
Upper bound	0.75	0.80	0.80	0.83	0.86	0.87	0.90	0.92	0.95
Osorio and Wang	0.76	0.78	0.83	0.83	0.84	0.87	0.89	0.90	0.94
Kerbache and Smith (1988)	0.64	0.72	0.66	0.74	0.79	0.80	0.83	0.86	0.92
Singh and Smith (1997)	0.72	0.78	0.74	0.81	0.85	0.86	0.88	0.90	0.95

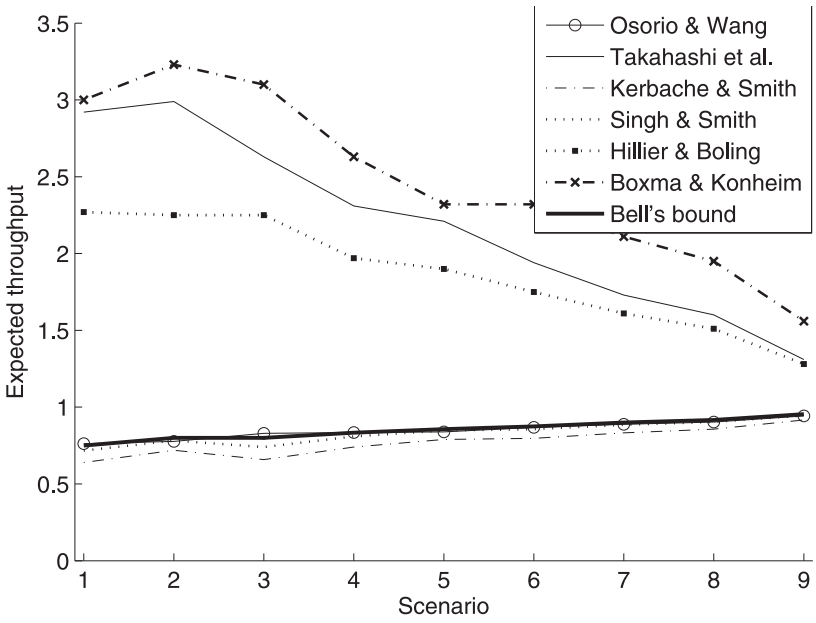

Fig. 22. Comparison of the expected throughput approximation of various decomposition methods with the theoretical upper bound derived by Bell (1982).

4. Case study

In this section, we use the proposed model to address a traditional urban traffic signal control problem. The purpose of this section is to investigate the added value of accounting for joint distributional information in traffic control. We compare the performance of the proposed approach with that of a similar decomposition approach that yields only univariate marginal (disaggregate) queue-length distributions. We call the latter approach the marginal queueing model. The comparison of the proposed approach with the marginal queueing model investigates the potential added value of accounting for multivariate joint distributions, i.e., multi-queue joint distributions, for traffic control. More generally, it investigates the potential added value of using models that provide a more detailed (i.e., beyond first-order) description of the between-queue interactions. This is of particular interest in congested urban networks, where vehicular spillbacks lead to the spatial propagation of congestion. For congested cities with short links and grid-type topologies, such as New York city (Osorio et al., 2014), the development of traffic management strategies that indeed mitigate vehicular spillbacks is critical. In order to design strategies that indeed limit spillbacks, models that account for a detailed description of between-queue dependencies are needed.

The formulation of the marginal queueing model for a multi-server general topology queueing network is derived in Osorio and Bierlaire (2009). Its formulation for a road network is presented in Osorio (2010, Chap. 4). The formulation for a single-server general topology queueing network is given in Appendix C of the present paper.

In order to map a road network as a queueing network, we follow the procedure presented in Osorio (2010, Chap. 4). We summarize it briefly here. Each lane in the road network is modeled as one (or multiple) queue(s). For the case study of this paper, all roads are single-lane roads and hence each road is mapped as a single queue. Each queue has a single server and finite space capacity defined by: $k_i = \lfloor (l_i + d_2)/(d_1 + d_2) \rfloor$, where l_i is the length of lane i in meters, d_1 is the average vehicle length (set to 4 m), and d_2 is the minimal inter-vehicle distance (set to 1 m). The fraction is rounded down to the nearest integer. This expression can be interpreted as decomposing the link into k_i slots, where the length of each slot corresponds to the average length a vehicle would occupy if the link were full of vehicles (i.e., under jam density traffic conditions). The routing probability matrix (p_{ij}) is obtained from link-to-link turning probabilities. External arrivals to link i arise following a Poisson process with rate parameter γ_i. These rates are obtained from an origin-destination matrix of the road network. The service rate of a queue corresponds to the downstream flow capacity of the underlying lane. For signalized lanes the service rate is defined by:

$$\mu_i = g_i s, \tag{35}$$

where g_i represents the total green split of queue i (i.e., ratio of total green time to intersection cycle time). Hence, in the signal control problem that we consider, a change in the green times leads to a change in the service rates of the corresponding lanes.

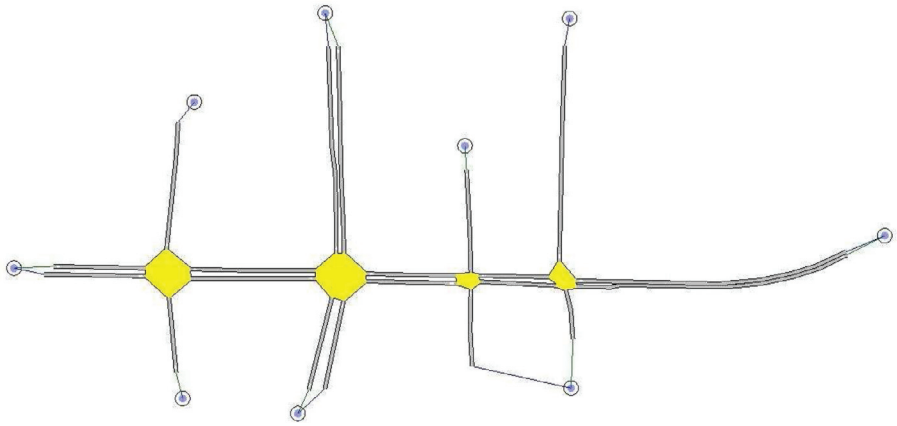

Fig. 23. Topology of urban network.

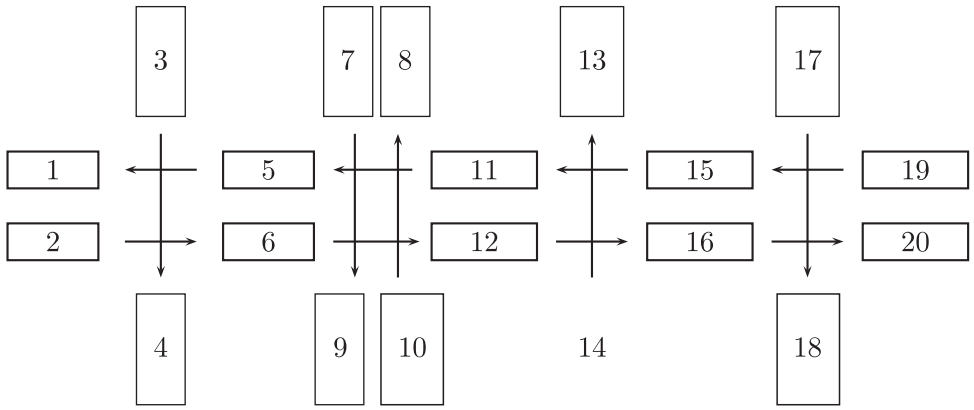

Fig. 24. Queueing network.

4.1. Network

We consider the road network displayed in Fig. 23. It considers 20 single-lane roads (i.e. 20 queues) and 4 intersections, each with 2 endogenous signal phases. Drivers travel along a single direction (i.e. they do not turn within the network). External arrivals and departures to the network occur at the boundaries of the network (represented by the circles in Fig. 23). This network has particularly simple traffic dynamics (e.g., no left or right turnings are allowed). The purpose of the case study is to show that even for such simple networks there is an added value of using joint distributional information for traffic control.

The representation of the road network as a queueing network is presented in Fig. 24. The queues are represented by rectangles, and labeled with indices 1–20. The arrows represent possible turning movements from one queue to another. For any pair of adjacent queues (i, j) connected by a straight arrow from i to j: p_{ij} equals 1, otherwise p_{ij} equals 0.

For the proposed model, the decomposition of the network into systems of queues is as follows. All west-bound links (i.e., roads) of the main arterial are modeled jointly, as are all east-bound links. The ross streets (north-bound and south-bound) are modeled individually (i.e. they are not part of a joint system) with the marginal model of Appendix C. In other words, the systems of the network are: (2, 6, 12), (6, 12, 16), (12, 16, 20), (11, 5, 1), (15, 11, 5), (19, 15, 11), (3), (4), (7), (8), (9), (10), (13), (14), (17), (18), where the numbers within parenthesis are queue indices.

We consider two different demand scenarios. For the medium demand scenario, the east-bound and west-bound demand (i.e. demand along the main arterial) is 700 vehicles per hour in each direction. This increases to 900 vehicles per hour in the high demand scenario. The demand is presented in Table 9, where the indices in the first row correspond to queue indices as defined in Fig. 24.

4.2. Problem formulation

The signal control problem that we consider is known as a fixed-time signal control problem. We briefly present its formulation here. For a review of traffic signal control terminology and formulations, we refer the reader to Appendix A

Table 9
Demand in vehicles per hour for the medium and high demand scenarios.

Demand scenario	19 → 1	2 → 20	3 → 4	7 → 9	10 → 8	14 → 13	17 → 18
Medium	700	700	100	600	600	100	100
High	900	900	100	600	600	200	200

of Osorio (2010). A fixed-time signal plan is also called time-of-day or pre-timed signal plan. These are strategies that use historical traffic patterns to derive a fixed signal plan for a given time period. The signal control problem is solved offline. The signal plans of multiple intersections are determined jointly. The decision variables are the green splits (i.e., normalized green times) of phases of the different intersections. All other traditional control variables (e.g., cycle times, offsets, stage structure) are assumed fixed.

Fixed-time signal plans are the most traditional form of signal timing. They do not rely on real-time traffic data. Hence, they are the standard practice in many cities with low, or inexistent, deployment of traffic sensors. For major cities with abundant real-time traffic data (e.g., New York City, Osorio et al. (2014)), fixed-time plans are used for time periods where congestion is both high and uniformly distributed. This occurs, for instance, along congested arterials with high levels of demand on the cross streets. The network of this paper is representative of such a situation. For congested networks with complex topologies (e.g., grid topologies), fixed-time plans are also commonly used. A variety of cities, including New York, use fixed-time plans to design traffic responsive plans. More specifically, a set of fixed-time plans are designed offline and are then selected in real-time based on prevailing traffic patterns.

To formulate this problem we introduce the following notation:

b_i	available cycle ratio of intersection i;
s	saturation flow rate [veh/h];
$x(j)$	green split of phase j;
x_L	vector of minimal green splits;
$\mathcal{I}$	set of intersection indices;
$\mathcal{L}$	set of indices of the signalized lanes;
$\mathcal{P}_I(i)$	set of phase indices of intersection i;
$\mathcal{P}_L(\ell)$	set of phase indices of lane ℓ.

The problem is formulated as follows:

$$\min_x T(x, y; u) \tag{36}$$

subject to

$$\sum_{j \in \mathcal{P}_I(i)} x(j) = b_i, \ \forall i \in \mathcal{I} \tag{37}$$

$$\mu_\ell - \sum_{j \in \mathcal{P}_L(\ell)} x_j s = 0, \ \forall \ell \in \mathcal{L} \tag{38}$$

$$h(y; u) = 0 \tag{39}$$

$$y \geq 0 \tag{40}$$

$$x \geq x_L, \tag{41}$$

where the decision vector x consists of the green splits for each phase. Constraints (37) ensure that for a given intersection the available cycle time is distributed among all phases. Constraint (38) relates the service rate (i.e., link flow capacity) of a signalized queue to the saturation flow s (set to 1800 vehicles per hour) and to its green split x_j. Eq. (39) represents the queueing model, i.e. the system of equations that is solved in order to yield the queue-length distributions, and the corresponding delays. The queueing model h depends on a vector of endogenous queueing variables y (e.g., disaggregation probabilities) and a set of exogenous parameters u (e.g., external arrival rates, space capacities). The endogenous queueing variables are subject to positivity constraints (40). Green splits have lower bounds (Eq. (41)), which are set to 4 s in this work (following the transportation norms VSS (1992)). The objective function $T(x, y; u)$ represents the expected trip travel time.

The expected time in the system is obtained by applying Little's law: (Little, 1961; 2011):

$$T(x, y; u) = \frac{\sum_i E[N_i]}{\sum_i \gamma_i P(N_i < k_i)}, \tag{42}$$

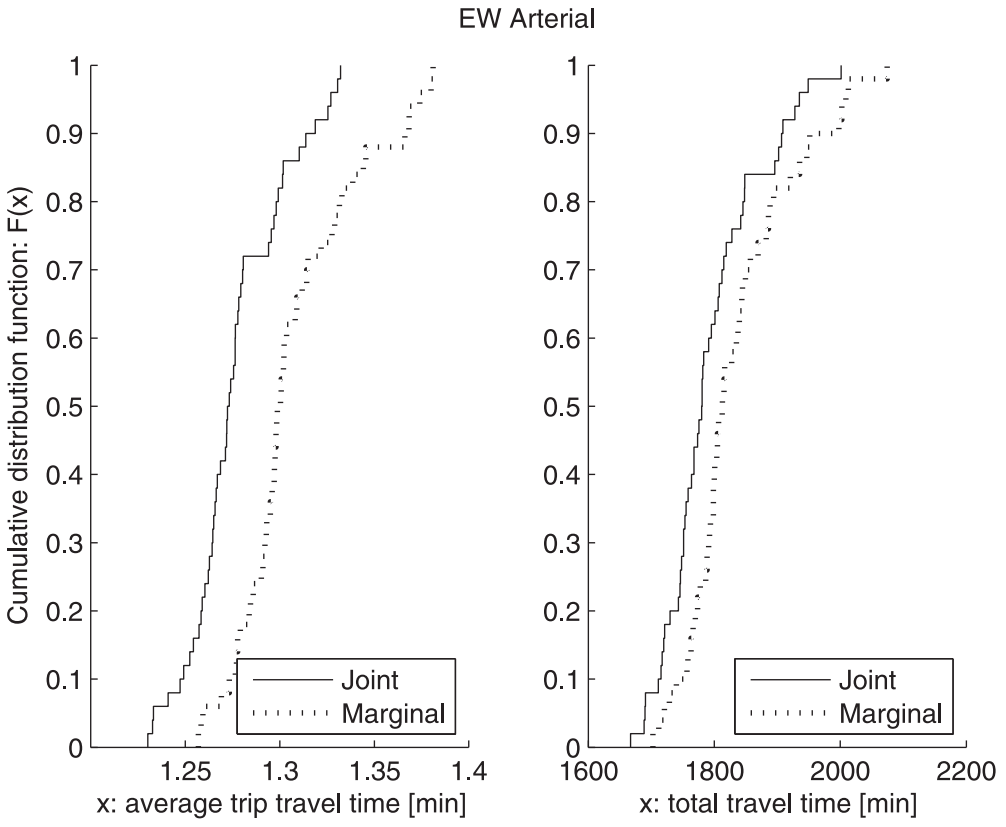

Fig. 25. Cdf's of the average (left plot) and total (right plot) trip travel time along the arterial for the medium demand scenario.

where $E[N_i]$ represents the expected number of vehicles in queue i and the summation considers all queues in the network (queues are indexed by i). The analytical expression for $E[N_i]$ is given in Appendix B.

4.3. Implementation notes

The marginal model is initialized as follows. For each intersection equal green splits are set for all endogenous phases (i.e. for phase j of intersection i: $x(j) = b_i/\text{card}(\mathcal{P}_I(i))$, where card is the cardinality function). Given the initial green splits, the remaining endogenous variables are obtained by solving the network model (i.e. solving the system of nonlinear Eqs. (50)). This is done with the "trust-region-dogleg" algorithm within the *fsolve* solver of Matlab, with constraint and function tolerances of 10^{-7}. This set of variables is used as an initial feasible point for the signal control problem, which is then solved using the "active-set" algorithm of the *fmincon* solver of Matlab with constraint and function tolerances of 10^{-6} and 10^{-3}, respectively.

The optimal signal plan proposed by the marginal model is then used as an initial signal plan for the joint model. The effective service rates $\hat{\mu}$ are initialized with the exogenous service rates μ, arrival rates λ are initialized by using the following variation of the flow conservation constraints:

$$\lambda_i = \gamma_i + \lambda_{i-1}. \tag{43}$$

Then initial marginal queue-length distributions are obtained by assuming the functional form in (14) along with $\rho = \lambda/\hat{\mu}$. The corresponding joint queue-length distributions are initialized by taking the products of the corresponding marginal distributions. This initial set of endogenous variables is used as an (infeasible) initial point for the signal control problem, which is solved with the sequential quadratic programming (SQP) algorithm of the *fmincon* solver with a constraint tolerance of 10^{-6} and a function tolerance of 10^{-3}.

In the signal control problem, we also implement the expected number of vehicles in each queue, $E[N_i]$, as a variable. Thus, in an I-queue network, there are I additional variables. The case study network (Fig. 23) consists of 20 single-lane roads which are modeled in the joint model as follows: two five-queue systems (i.e., two sets of 5 lanes that are modeled jointly) and 10 lanes are modeled individually (i.e., they are not a part of a joint system). Each five-queue system leads to $41(5-2)+2(5) = 133$ variables and $41(5-2)+2(5)+9(5-3) = 151$ equations. Each individually modeled lane leads to a set of 5 variables and 5 constraints. The joint model therefore consists of $2(133)+50 = 316$ variables and $2(151)+50 = 352$ constraints.

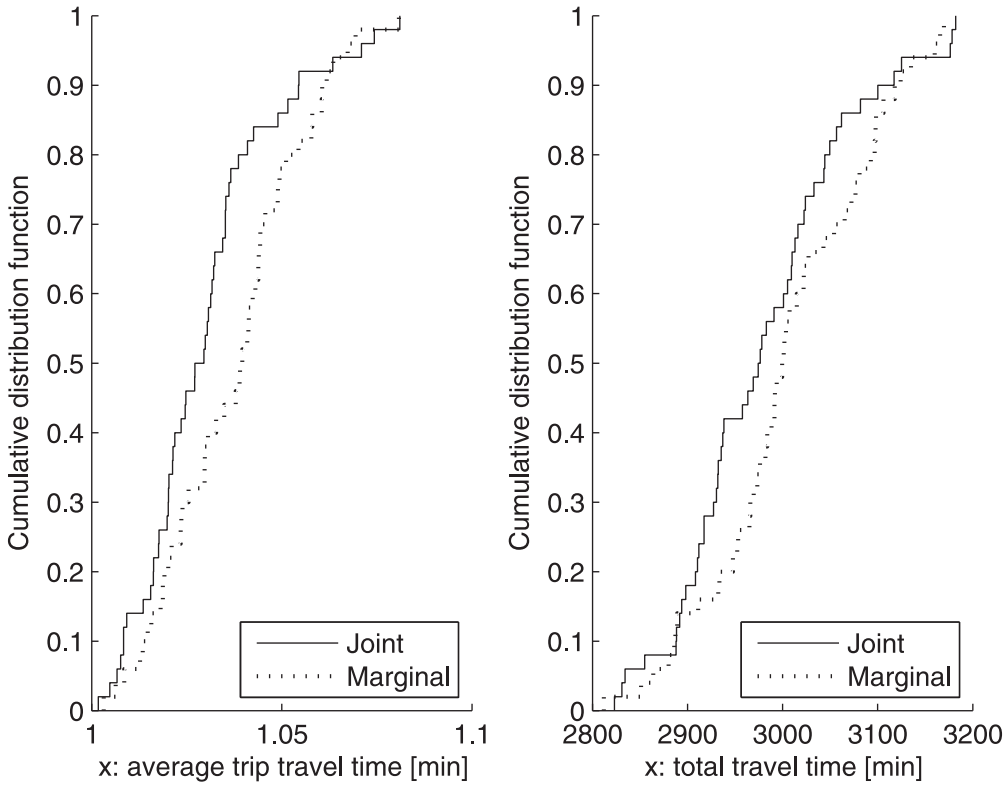

Fig. 26. Cdf's of the average (left plot) and total (right plot) trip travel time for the full network and the medium demand scenario.

The signal control problem has the following additional endogenous variables and constraints: 8 phase variables (2 per intersection) with their corresponding lower bound constraints (Eq. (41)), 4 green split allocation linear constraints (Eq. (37)) (1 per intersection). Thus, the signal control system consists of 324 variables, 356 equality constraints, and 8 inequality constraints.

4.4. Results

The performance of the signal plans proposed by both the marginal and the joint formulations are evaluated by a microscopic traffic simulation model implemented in Aimsun version 6.1 (TSS, 2011). For a given signal plan, we embed it within the simulator and run 50 simulation replications, each for 1 h with a warm-up period of 15 min.. For each replication, we obtain a given simulated performance measure (e.g., average trip travel time). We compare the cumulative distribution functions (cdf's) obtained from the 50 observations of the performance measure.

4.4.1. Medium demand scenario

We first study the performance of the links on the main arterial (i.e., east-bound and west-bound). These are the links that are modeled jointly under the proposed approach (i.e., their joint distribution is approximated). Fig. 25 considers two plots. The left (resp. right) plot displays the cdf of the average (resp. total) trip travel times for the trips along the main arterial. Each plot displays two cdf curves, the solid curve corresponds to the proposed joint model, the dashed curve corresponds to the marginal model. Let us detail how to interpret a cdf curve. The x-axis displays the performance measure (either average or total trip travel time). For a given x value the y-axis displays the proportion of simulation replications (out of the 50) where the simulated (average or total) trip travel time is smaller than x. Hence, the more the cdf curve is shifted to the left, the higher the proportion of simulation replications with smaller (average or total) trip travel time values.

Fig. 25 displays two plots. In both plots of Fig. 25, the cdf curve of the signal plan proposed by the joint model is to the left of that of the marginal model. Hence, the signal plan proposed by the joint model leads to lower average trip travel times, and lower total travel times, compared to the signal plan proposed by the marginal model. The joint model yields a signal plan with a 2.4% improvement in the average trip travel time.

We test the hypothesis that the expected arterial trip travel time derived from the joint model is equal to the time derived by the marginal model for the medium demand scenario by conducting a paired t-test. We consider a one-sided test, the null hypothesis assumes that the expectation is equal under both signal plans, the alternative hypothesis assumes that the expectation of the signal plan derived by the joint model is smaller than that derived by the marginal model. We

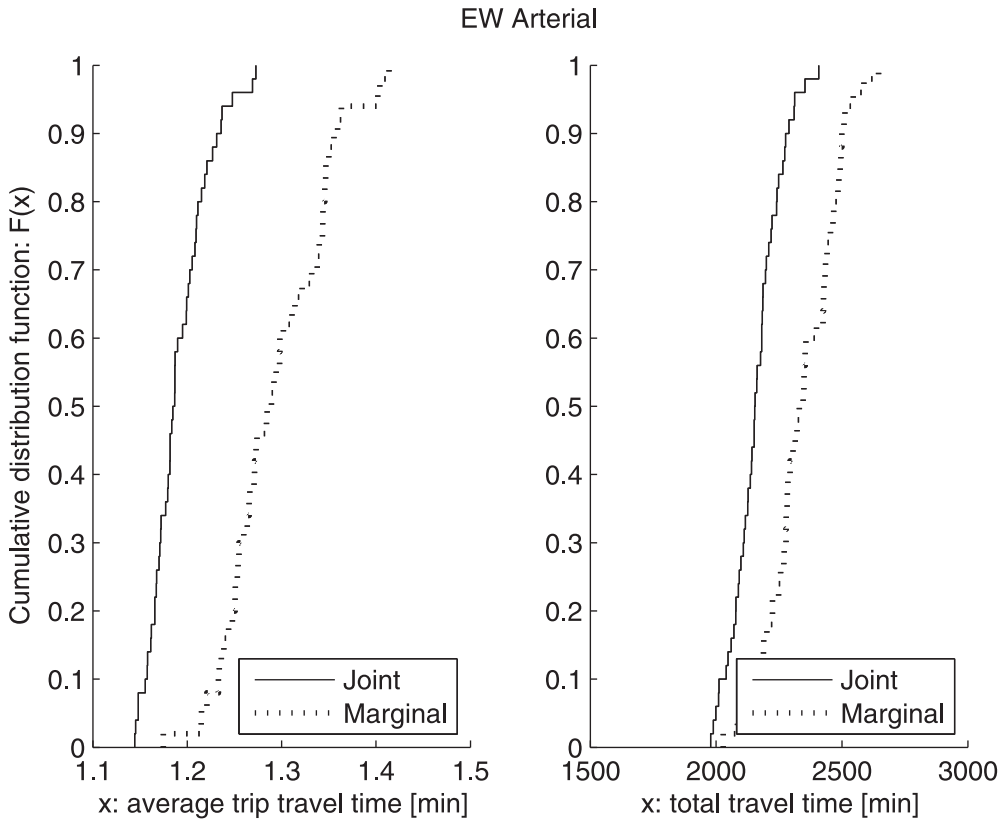

Fig. 27. Cdf's of the average (left plot) and total (right plot) trip travel time along the arterial for the high demand scenario.

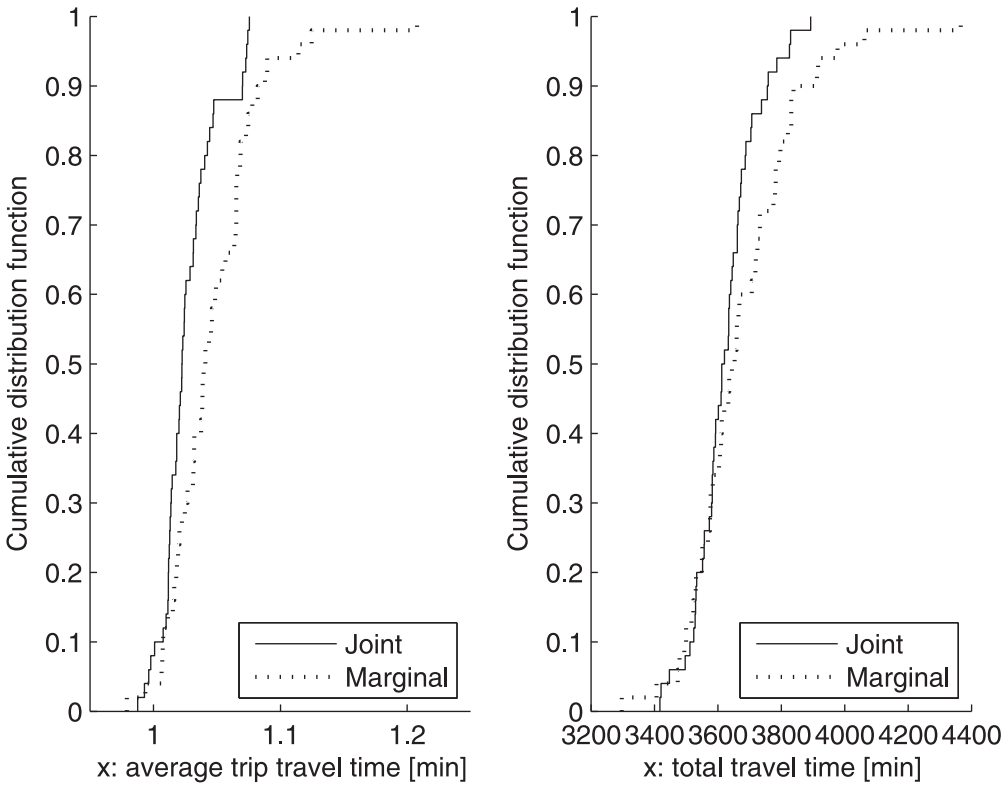

Fig. 28. Cdf's of the average (left plot) and total (right plot) trip travel time for the full network and the high demand scenario.

consider a significance level of 5%. The test has 49 degrees of freedom, and the critical value is -1.677. The test statistic is −19.4, which is smaller than the critical value. Hence, the null hypothesis of equal expectation is rejected. This means that the expected trip travel time derived by the signal plan of the joint model is statistically significantly lower than that of the signal plan of the marginal model.

Fig. 26 considers all links in the network (i.e., both the arterial links and the cross streets). In both plots of Fig. 26, the cdf curve of the signal plan proposed by the joint model is to the left of that of the marginal model. Hence, the signal plan proposed by the joint model leads to lower average trip travel times, and lower total travel times, compared to the signal plan proposed by the marginal model.

4.4.2. High demand scenario

The high demand scenario increases the demand relative to the medium demand scenario along the main arterial and two of the cross streets. We proceed in the analysis as for the medium demand scenario. Fig. 27 considers the performance of the links on the main arterial (i.e., east-bound and west-bound). These are the links that are modeled jointly under the proposed approach. The left (resp. right) plot of Fig. 27 displays the cdf of the average (resp. total) trip travel times for the trips along the main arterial. Just as for the medium demand scenario, the signal plan proposed by the joint model leads to average trip travel times and total trip travel times, that are both lower than those obtained from the signal plan of the marginal model. The joint model yields a signal plan with a 7.8% improvement in the average trip travel time. We carry out the same paired one-sided t-test as we did for the medium demand scenario. The test statistic is -19.6, which is smaller than the critical value of -1.677. Hence, the null hypothesis of equal expectation is rejected. The joint model yields significantly smaller expected trip travel times along the arterial than the marginal model.

Fig. 28 considers all links in the network (i.e., both arterial links and cross streets), it displays the same performance measures as Fig. 27. Here again, the cdf's of the joint model are to the left of those of the marginal model, i.e., they yield lower average trip travel times and lower total trip travel times.

For both the medium and the high demand scenario, there is a statistically significant improvement in the travel times of the links that are modeled jointly. This illustrates the added value of accounting for a more detailed (i.e., beyond first-order) description of between-link dependency. These results also indicate that this information is of added value for both medium and high levels of congestion.

5. Conclusion

This paper proposes an analytical approximation of the stationary aggregate joint queue-length distribution of a tandem Markovian network. The method combines ideas from decomposition methods, finite capacity queueing network models and aggregation-disaggregation techniques. The model consists of a system of nonlinear equations with a dimension that increases linearly with the number of queues, rather than exponentially, and that is independent of the space capacity of the individual queues. This makes the model suitable for the analysis of large-scale networks. The analytical joint distribution is validated versus simulation estimates and versus other decomposition methods. This queueing method is then used to model a congested urban traffic network, and to address a traditional signal control problem. The problem is solved with the proposed joint modeling approach and with an analytical model that only approximates univariate marginal distributions, i.e., it only captures first-order between-queue dependency information. The proposed model yields signal plans with significantly lower average trip travel times and lower total travel times. This case study illustrates the added value of using higher-order spatial dependency information for traffic control.

Ongoing work has extended these stationary models to time-dependent models. More specifically, we have formulated an analytical and tractable approximation of the aggregate joint transient queue-length distribution (Osorio and Yamani, 2016). The latter work illustrates the added value of using a model that accounts for both joint and transient information for signal control. Compared to the results of this paper, the approach of Osorio and Yamani (2016) outperforms the present approach, yet at the cost of reduced computational efficiency. In particular, the stationarity assumption of the proposed approach has lead to a tractable formulation defined as a system of nonlinear equations. The Osorio and Yamani (2016) method provides a temporal description of the joint distribution, yet its evaluation requires the solution of a series of systems of equations, as well as a matrix exponential computation, and this at every time step of the algorithm. Ongoing work is exploring ideas to further enhance the computational efficiency of these methods. The extension of the proposed stationary model to account for arbitrary topology networks, as well as its use to enhance the computational efficiency of simulation-based optimization methods (Osorio and Bierlaire, 2013; Osorio and Chong, 2015) is of interest.

Acknowledgments

This research was partially supported by the Center for Complex Engineering Systems at KACST and MIT. 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.

Appendix A. Transition rate matrix for a three-queue system

Table 10 details the full transition rate matrix of a three queue system with queues indexed by $(i, i+1, i+2)$ within a general I-queue network. The parameters μ, γ and k are exogenous. The variables λ_i and $\hat{\mu}_{i+2}$ are endogenous. Recall that λ_i approximates the arrivals to queue i, which consist of either external arrivals, γ_i, or arrivals from queues upstream of i which are not in the $(i, i+1, i+2)$ three-queue system. Similarly, the service rate of queue $i+2$ (the most downstream of the three-queue system), $\hat{\mu}_{i+2}$, is also endogenous, since it is determined by the traffic at further downstream queues.

External arrivals are allowed to all queues. The first three sets of rates define the transitions that may occur when an external arrival occurs at queue i, $i+1$ and $i+2$, respectively. An external arrival to a given queue can cause the queue to transition from aggregate state 0 (resp. 1) to aggregate state 1 (resp. 2). Upon an external arrival to, for instance, queue i, the transition from aggregate state 0 to 1 occurs with probability 1 and the transition from aggregate state 1 to 2 occurs if queue i is in the disaggregate state $k_i - 1$. This occurs with probability $\alpha^{\mathrm{f}}_{i,s}$ (defined in Section 2.3.1).

The fourth set of rates considers a service completion at queue i. Such an event can cause queue i to transition from aggregate state 1 (resp. 2) to 0 (resp. 1). Upon service completion, the transition from state 2 to 1 occurs with probability 1 and the transition from 1 to 0 occurs if queue i is in the disaggregate state 1. This occurs with probability $\alpha^{\mathrm{e}}_{i,s}$ (defined in Section 2.3.1). Additionally, a service completion at queue i can cause queue $i+1$ to transition from aggregate state 0 (resp. 1) to 1 (resp. 2), which occurs with probability 1 (resp. $\alpha^{\mathrm{f}}_{i+1,s}$).

The fifth set of rates considers a service completion at queue $i+1$. The rates are obtained through similar reasoning as for service completion at queue i. Additionally, if a job at queue i is being blocked by queue $i+1$, then a service completion at queue $i+1$ may trigger a change in the state of queue i. This is described via the blocking probability $\beta_{i,1}$ (defined in Section 2.3.2). More specifically, if queue i is blocked by queue $i+1$, then a service completion at queue $i+1$ will:

1. send the job that has completed service at queue $i+1$ to queue $i+2$, which may lead queue $i+2$ to transition from aggregate state 0 (resp. 1) to 1 (resp. 2);
2. unblock a job at queue i, which may lead queue i to transition from aggregate state 1 (resp. 2) to 0 (resp. 1);
3. have no impact on the state of queue $i+1$ (since an arrival and a departure occur simultaneously).

The final set of rates considers a service completion at queue $i+2$. The rates are obtained through similar reasoning as for service completion at queue $i+1$. Since queue $i+2$ may block both queue $i+1$ and queue i, then a service completion at queue $i+2$ may trigger changes in the states of both queues i and $i+1$. This unblocking is described via the blocking probabilities $\beta_{i,2}$, $\beta_{i,3}$ and $\beta_{i,4}$ (defined in Section 2.3.2).

Appendix B. Expected number of vehicles

This section derives the analytical expression for the expected number of vehicles in queue i, $E[N_i]$. We have:

$$\begin{aligned} E[N_i] &= E[E[N_i \mid N_{A,i}]] \\ &= 0P(N_{A,i}=0) + E[N_i \mid N_{A,i}=1]P(N_{A,i}=1) + k_i P(N_{A,i}=2) \end{aligned} \tag{44}$$

We derive an analytical approximation for $E[N_i \mid N_{A,i}=1]$. By definition:

$$E[N_i \mid N_{A,i}=1] = \sum_{n=1}^{k_i-1} nP(N_i = n \mid N_{A,i}=1) \tag{45}$$

We approximate $P(N_i \mid N_{A,i}=1)$ by using the functional form for the stationary distribution of a single M/M/1/k queue (Eq. (14)), and following a similar derivation to that of (15). For the traffic intensity of link segment i, ρ_i, we obtain:

$$P(N_i = n \mid N_{A,i}=1) = \frac{(1-\rho_i)\rho_i^{n-1}}{1-\rho_i^{k_i-1}}. \tag{46}$$

Inserting (46) into (45):

$$\begin{aligned} E[N_i \mid N_{A,i}=1] &= \sum_{n=1}^{k_i-1} n\frac{(1-\rho_i)\rho_i^{n-1}}{1-\rho_i^{k_i-1}} = \frac{1-\rho_i}{1-\rho_i^{k_i-1}}\sum_{n=1}^{k_i-1} n\rho_i^{n-1} \\ &= \frac{1-\rho_i}{1-\rho_i^{k_i-1}}\sum_{n=1}^{k_i-1}\frac{d}{d\rho_i}\rho_i^n = \frac{1-\rho_i}{1-\rho_i^{k_i-1}}\frac{d}{d\rho_i}\left(\sum_{n=1}^{k_i-1}\rho_i^n\right) \\ &= \frac{1-\rho_i}{1-\rho_i^{k_i-1}}\frac{d}{d\rho_i}\left(\rho_i\sum_{n=0}^{k_i-2}\rho_i^n\right) \\ &= \frac{1-\rho_i}{1-\rho_i^{k_i-1}}\frac{d}{d\rho_i}\left(\rho_i\frac{1-\rho_i^{k_i-1}}{1-\rho_i}\right) \end{aligned}$$

Table 10
Transition rate matrix for a three-queue system with the 3 queues indexed by $(i, i+1, i+2)$. Enumeration of all possible transitions assuming an initial state $s = (j_i, j_{i+1}, j_{i+2})$ and a new state t.

External arrival to queue i

New state t	Initial conditions	Rate
$(j_i+1, j_{i+1}, j_{i+2})$	$j_i = 0$	λ_i
$(j_i+1, j_{i+1}, j_{i+2})$	$j_i = 1,\ j_{i+1} = \{0,1\}$	$\lambda_i \alpha^{\mathrm{f}}_{i,s}$
$(j_i+1, j_{i+1}, j_{i+2})$	$j_i = 1,\ j_{i+1} = 2,\ j_{i+2} = \{0,1\}$	$\lambda_i \alpha^{\mathrm{f}}_{i,s}$
$(j_i+1, j_{i+1}, j_{i+2})$	$j_i = 1,\ j_{i+1} = 2,\ j_{i+2} = 2$	$\lambda_i \alpha^{\mathrm{f}}_{i,s}$

External arrival to queue $i+1$

New state t	Initial conditions	Rate
$(j_i, j_{i+1}+1, j_{i+2})$	$j_{i+1} = 0$	γ_{i+1}
$(j_i, j_{i+1}+1, j_{i+2})$	$j_{i+1} = 1,\ j_{i+2} = \{0,1\}$	$\gamma_{i+1} \alpha^{\mathrm{f}}_{i+1,s}$
$(j_i, j_{i+1}+1, j_{i+2})$	$j_{i+1} = 1,\ j_{i+2} = 2$	$\gamma_{i+1} \alpha^{\mathrm{f}}_{i+1,s}$

External arrival to queue $i+2$

New state t	Initial conditions	Rate
$(j_i, j_{i+1}, j_{i+2}+1)$	$j_{i+2} = 0$	γ_{i+2}
$(j_i, j_{i+1}, j_{i+2}+1)$	$j_{i+2} = 1$	$\gamma_{i+2} \alpha^{\mathrm{f}}_{i+2,s}$

Service completion at queue i

New state t	Initial conditions	Rate
$(j_i, j_{i+1}+1, j_{i+2})$	$j_i = 1,\ j_{i+1} = 0$	$\mu_i(1-\alpha^{\mathrm{e}}_{i,s})$
$(j_i-1, j_{i+1}+1, j_{i+2})$	$j_i = 1,\ j_{i+1} = 0$	$\mu_i \alpha^{\mathrm{e}}_{i,s}$
$(j_i, j_{i+1}+1, j_{i+2})$	$j_i = 1,\ j_{i+1} = 1,\ j_{i+2} = \{0,1\}$	$\mu_i(1-\alpha^{\mathrm{e}}_{i,s})\alpha^{\mathrm{f}}_{i+1,s}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i = 1,\ j_{i+1} = 1,\ j_{i+2} = \{0,1\}$	$\mu_i \alpha^{\mathrm{e}}_{i,s}(1-\alpha^{\mathrm{f}}_{i+1,s})$
$(j_i-1, j_{i+1}+1, j_{i+2})$	$j_i = 1,\ j_{i+1} = 1,\ j_{i+2} = \{0,1\}$	$\mu_i \alpha^{\mathrm{e}}_{i,s}\alpha^{\mathrm{f}}_{i+1,s}$
$(j_i, j_{i+1}+1, j_{i+2})$	$j_i = 1,\ j_{i+1} = 1,\ j_{i+2} = 2$	$\mu_i(1-\alpha^{\mathrm{e}}_{i,s})\alpha^{\mathrm{f}}_{i+1,s}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i = 1,\ j_{i+1} = 1,\ j_{i+2} = 2$	$\mu_i \alpha^{\mathrm{e}}_{i,s}(1-\alpha^{\mathrm{f}}_{i+1,s})$
$(j_i-1, j_{i+1}+1, j_{i+2})$	$j_i = 1,\ j_{i+1} = 1,\ j_{i+2} = 2$	$\mu_i \alpha^{\mathrm{e}}_{i,s}\alpha^{\mathrm{f}}_{i+1,s}$
$(j_i-1, j_{i+1}+1, j_{i+2})$	$j_i = 2,\ j_{i+1} = 0$	μ_i
$(j_i-1, j_{i+1}+1, j_{i+2})$	$j_i = 2,\ j_{i+1} = 1,\ j_{i+2} = \{0,1\}$	$\mu_i \alpha^{\mathrm{f}}_{i+1,s}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i = 2,\ j_{i+1} = 1,\ j_{i+2} = \{0,1\}$	$\mu_i(1-\alpha^{\mathrm{f}}_{i+1,s})$
$(j_i-1, j_{i+1}+1, j_{i+2})$	$j_i = 2,\ j_{i+1} = 1,\ j_{i+2} = 2$	$\mu_i \alpha^{\mathrm{f}}_{i+1,s}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i = 2,\ j_{i+1} = 1,\ j_{i+2} = 2$	$\mu_i(1-\alpha^{\mathrm{f}}_{i+1,s})$

Service completion at queue $i+1$

New state t	Initial conditions	Rate
$(j_i, j_{i+1}, j_{i+2}+1)$	$j_{i+1} = 1,\ j_{i+2} = 0$	$\mu_{i+1}(1-\alpha^{\mathrm{e}}_{i+1,s})$
$(j_i, j_{i+1}-1, j_{i+2}+1)$	$j_{i+1} = 1,\ j_{i+2} = 0$	$\mu_{i+1}\alpha^{\mathrm{e}}_{i+1,s}$
$(j_i, j_{i+1}, j_{i+2}+1)$	$j_{i+1} = 1,\ j_{i+2} = 1$	$\mu_{i+1}(1-\alpha^{\mathrm{e}}_{i+1,s})\alpha^{\mathrm{f}}_{i+2,s}$
$(j_i, j_{i+1}-1, j_{i+2})$	$j_{i+1} = 1,\ j_{i+2} = 1$	$\mu_{i+1}\alpha^{\mathrm{e}}_{i+1,s}(1-\alpha^{\mathrm{f}}_{i+2,s})$
$(j_i, j_{i+1}-1, j_{i+2}+1)$	$j_{i+1} = 1,\ j_{i+2} = 1$	$\mu_{i+1}\alpha^{\mathrm{e}}_{i+1,s}\alpha^{\mathrm{f}}_{i+2,s}$
$(j_i, j_{i+1}-1, j_{i+2}+1)$	$j_i = 0,\ j_{i+1} = 2,\ j_{i+2} = 0$	μ_{i+1}
$(j_i, j_{i+1}-1, j_{i+2})$	$j_i = 0,\ j_{i+1} = 2,\ j_{i+2} = 1$	$\mu_{i+1}(1-\alpha^{\mathrm{f}}_{i+2,s})$
$(j_i, j_{i+1}-1, j_{i+2}+1)$	$j_i = 0,\ j_{i+1} = 2,\ j_{i+2} = 1$	$\mu_{i+1}\alpha^{\mathrm{f}}_{i+2,s}$
$(j_i, j_{i+1}-1, j_{i+2}+1)$	$j_i = \{1,2\},\ j_{i+1} = 2,\ j_{i+2} = 0$	$\mu_{i+1}(1-\beta_{i,1})$
$(j_i, j_{i+1}-1, j_{i+2})$	$j_i = \{1,2\},\ j_{i+1} = 2,\ j_{i+2} = 1$	$\mu_{i+1}(1-\alpha^{\mathrm{f}}_{i+2,s})(1-\beta_{i,1})$
$(j_i, j_{i+1}-1, j_{i+2}+1)$	$j_i = \{1,2\},\ j_{i+1} = 2,\ j_{i+2} = 1$	$\mu_{i+1}\alpha^{\mathrm{f}}_{i+2,s}(1-\beta_{i,1})$
$(j_i, j_{i+1}, j_{i+2}+1)$	$j_i = 1,\ j_{i+1} = 2,\ j_{i+2} = 0$	$\mu_{i+1}(1-\alpha^{\mathrm{e}}_{i,s})\beta_{i,1}$
$(j_i-1, j_{i+1}, j_{i+2}+1)$	$j_i = 1,\ j_{i+1} = 2,\ j_{i+2} = 0$	$\mu_{i+1}\alpha^{\mathrm{e}}_{i,s}\beta_{i,1}$
$(j_i, j_{i+1}, j_{i+2}+1)$	$j_i = 1,\ j_{i+1} = 2,\ j_{i+2} = 1$	$\mu_{i+1}(1-\alpha^{\mathrm{e}}_{i,s})\alpha^{\mathrm{f}}_{i+2,s}\beta_{i,1}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i = 1,\ j_{i+1} = 2,\ j_{i+2} = 1$	$\mu_{i+1}(1-\alpha^{\mathrm{f}}_{i+2,s})\alpha^{\mathrm{e}}_{i,s}\beta_{i,1}$
$(j_i-1, j_{i+1}, j_{i+2}+1)$	$j_i = 1,\ j_{i+1} = 2,\ j_{i+2} = 1$	$\mu_{i+1}\alpha^{\mathrm{f}}_{i+2,s}\alpha^{\mathrm{e}}_{i,s}\beta_{i,1}$
$(j_i-1, j_{i+1}, j_{i+2}+1)$	$j_i = 2,\ j_{i+1} = 2,\ j_{i+2} = 0$	$\mu_{i+1}\beta_{i,1}$
$(j_i-1, j_{i+1}, j_{i+2}+1)$	$j_i = 2,\ j_{i+1} = 2,\ j_{i+2} = 1$	$\mu_{i+1}\alpha^{\mathrm{f}}_{i+2,s}\beta_{i,1}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i = 2,\ j_{i+1} = 2,\ j_{i+2} = 1$	$\mu_{i+1}(1-\alpha^{\mathrm{f}}_{i+2,s})\beta_{i,1}$

Service completion at queue $i+2$

New state t	Initial conditions	Rate
$(j_i, j_{i+1}, j_{i+2}-1)$	$j_{i+2} = 1$	$\hat{\mu}_{i+2}\alpha^{\mathrm{e}}_{i+2,s}$
$(j_i, j_{i+1}, j_{i+2}-1)$	$j_{i+1} = 0,\ j_{i+2} = 2$	$\hat{\mu}_{i+2}$
$(j_i, j_{i+1}, j_{i+2}-1)$	$j_{i+1} = \{1,2\},\ j_{i+2} = 2$	$\hat{\mu}_{i+2}(1-\beta_{i,3})$
$(j_i, j_{i+1}-1, j_{i+2})$	$j_{i+1} = 1,\ j_{i+2} = 2$	$\hat{\mu}_{i+2}\alpha^{\mathrm{e}}_{i+1,s}\beta_{i,3}$
$(j_i, j_{i+1}-1, j_{i+2})$	$j_i = 0,\ j_{i+1} = 2,\ j_{i+2} = 2$	$\hat{\mu}_{i+2}\beta_{i,3}$
$(j_i, j_{i+1}-1, j_{i+2})$	$j_i = \{1,2\},\ j_{i+1} = 2,\ j_{i+2} = 2$	$\hat{\mu}_{i+2}\beta_{i,4}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i = 1,\ j_{i+1} = 2,\ j_{i+2} = 2$	$\hat{\mu}_{i+2}\alpha^{\mathrm{e}}_{i,s}\beta_{i,2}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i = 2,\ j_{i+1} = 2,\ j_{i+2} = 2$	$\hat{\mu}_{i+2}\beta_{i,2}$

Table 11
List of variables used in marginal finite capacity queueing model.

γ_i	external arrival rate;
λ_i	total arrival rate;
μ_i	service rate of a server;
$\tilde{\mu}_i$	unblocking rate;
$\hat{\mu}_i$	effective service rate;
P_i^f	probability of being blocked at queue i;
p_{ij}	transition probability from queue i to queue j;
k_i	space capacity;
N_i	number of vehicles in queue i;
$P(N_i = k_i)$	probability that queue i is full;
$\mathcal{I}_i^+$	set of downstream queues to queue i.

$$= \frac{1-\rho_i}{1-\rho_i^{k_i-1}}\left(\frac{1-\rho_i^{k_i-1}}{1-\rho_i} + \rho_i\left[\frac{1-\rho_i^{k_i-1}}{(1-\rho_i)^2} - \frac{(k_i-1)\rho_i^{k_i-2}}{1-\rho_i}\right]\right)$$
$$= 1 + \frac{\rho_i}{1-\rho_i} - \frac{(k_i-1)\rho_i^{k_i-1}}{1-\rho_i^{k_i-1}}. \tag{47}$$

We proceed as in Section 2.3.1 and provide the following state-dependent approximation of ρ_i, for a state $s = (1, j_{i+1}, j_{i+2})$:

$$\rho_{i,s} = \begin{cases} \lambda_i/\mu_i & \text{if} \quad j_{i+1} < 2 \\ \lambda_i/\mu_{i+1} & \text{if} \quad j_{i+1} = 2 \quad \text{and} \quad j_{i+2} < 2 \\ \lambda_i/\hat{\mu}_{i+2} & \text{if} \quad j_{i+1} = 2 \quad \text{and} \quad j_{i+2} = 2, \end{cases} \tag{48}$$

where λ_i is obtained through (20), μ is exogenous and $\hat{\mu}$ is obtained through (30). To summarize, for a given queue i, we consider the three-queue system where i is the most upstream of the 3 queues. This allows us to derive the value of the disaggregate state N_i by conditioning on the states of the 2-downstream queues. In this way, we account for blocking that arises from either of these 2 downstream queues. Let P_i denote the aggregate joint distribution of the system $(N_{A,i}, N_{A,i+1}, N_{A,i+2})$.

$$\begin{aligned} E[N_i \mid N_{A,i} = 1] = E[E[N_i \mid N_{A,i} = 1, N_A = s]] &= E[N_i \mid N_{A,i} = 1, N_{A,i+1} < 2]P_i(N_{A,i} = 1, N_{A,i+1} < 2) \\ &+ E[N_i \mid N_{A,i} = 1, N_{A,i+1} = 2, N_{A,i+2} < 2]P_i(N_{A,i} = 1, N_{A,i+1} = 2, N_{A,i+2} < 2) \\ &+ E[N_i \mid N_{A,i} = 1, N_{A,i+1} = 2, N_{A,i+2} = 2]P_i(N_{A,i} = 1, N_{A,i+1} = 2, N_{A,i+2} = 2), \end{aligned} \tag{49}$$

where the three expectations on the right-hand side of the above equation are given by inserting the expression in (47) and their corresponding ρ values defined by (48).

Appendix C. Marginal finite capacity queueing model

The finite capacity queueing model of (Osorio, 2010, Chap. 4) is given by the System of Eq. (50). It approximates the marginal distribution of each queue. It accounts only for first-order between-queue dependency information. We refer to this model as the marginal finite capacity queueing model. The variables used in this model for a given queue i are listed in Table 11.

$$\begin{cases} \lambda_i = \gamma_i + \frac{\sum_j p_{ij}\lambda_j P(N_j<k_j)}{P(N_i<k_i)} & \text{(a)} \\ \frac{1}{\tilde{\mu}_i} = \sum_{j\in\mathcal{I}_i^+} \frac{\lambda_j P(N_j<k_j)}{\lambda_i P(N_i<k_i)\hat{\mu}_j} & \text{(b)} \\ \frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + P_i^f \frac{1}{\tilde{\mu}_i} & \text{(c)} \\ P_i^f = \sum_j p_{ij} P(N_j = k_j) & \text{(d)} \\ P(N_i = k_i) = \frac{1-\rho_i}{1-\rho_i^{k_i+1}}\rho^{k_i} & \text{(e)} \\ \rho_i = \frac{\lambda_i}{\hat{\mu}_i}. & \text{(f)} \end{cases} \tag{50}$$

The exogenous parameters are γ_i, p_{ij}, k_i, and μ_i. All other variables are endogenous. Eq. (50a) is a flow conservation equation as applied to a loss (finite capacity) queueing model. Eq. (50b) defines the unblocking rate of a queue that is blocked, (50c) defines the effective service rate, which accounts for both (exogenous) service, μ_i, and blocking, $\tilde{\mu}_i$. Eq. (50d) approximates the probability of being blocked at queue i by averaging the probabilities of downstream queues being full (which are called blocking probabilities). The blocking probability of queue i is given in Eq. (50e) by the closed-form expression of the stationary queue-length distribution of an M/M/1/k queue (see, e.g., Bocharov et al., 2004). Eq. (50f) defines the traffic intensity, ρ_i, of a finite capacity queue.

References

Alfa, A.S., Liu, B., 2004. Performance analysis of a mobile communication network: the tandem case. Comp. Comm. 27 (3), 208–221.

Altiok, T., 1982. Approximate analysis of exponential tandem queues with blocking. Eur. J. Oper. Res. 11 (4), 390–398.

Altiok, T., Perros, H.G., 1987. Approximate analysis of arbitrary configurations of open queuing networks with blocking. Ann. Oper. Res. 9 (1), 481–509.

Balsamo, S., De Nitto Persone, V., Onvural, R., 2001. Analysis of Queueing Networks with Blocking. International Series in Operations Research and Management Science, 31. Kluwer Academic Publishers, Boston.

Baskett, F., Chandy, K.M., Muntz, R., Palacios, F., 1975. Open, closed and mixed networks of queues with different classes of customers. J. Assoc. Comput. Mach. 22 (2), 248–260.

Bell, P.C., 1982. Use of decomposition techniques for the analysis of open restricted queuing networks. Oper. Res. Lett. 1 (6), 230–235.

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

Boxma, O.J., Konheim, A.J., 1981. Approximate analysis of exponential queueing systems with blocking. Acta Inform. 15 (1), 19–66.

Brandwajn, A., Jow, Y., 1985. Tandem exponential queues with finite buffers. In: Hasegawa, T., Takagi, H., Takahashi, Y. (Eds.), Comp. Networking and Perf. Evaluation. North Holland, Amsterdam, The Netherlands, pp. 245–258.

Brandwajn, A., Jow, Y., 1988. An approximation method for tandem queues with blocking. Oper. Res. 36 (1), 73–83.

Cheah, J.Y., Smith, J.M., 1994. Generalized M/G/C/C state dependent queueing models and pedestrian traffic flows. Queueing Syst. 15 (1–4), 365–386.

Chong, L., Osorio, C., 2016. A simulation-based optimization algorithm for dynamic large-scale urban transportation problems. Trans. Sci.. Forthcoming. Available at: http://web.mit.edu/osorioc/www/papers/osoChoDynSOsubmitted.pdf.

Dallery, Y., Gershwin, S.B., 1992. Manufacturing flow line systems: a review of models and analytical results. Queueing Syst. 12 (1–2), 3–94.

Department of Transportation, 2008. Transportation Vision for 2030. Technical Report. U.S. Department of Transportation (DOT), Research and Innovative Technology Administration.

Gershwin, S.B., 1987. An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking. Oper. Res. 35 (2), 291–305.

Gershwin, S.B., Burman, M.H., 2000. A decomposition method for analyzing inhomogenous assembly/disassembly systems. Ann. Oper. Res. 93, 91–115.

Gupta, S.M., Kavusturucu, A., 2000. Production systems with interruptions, arbitrary topology and finite buffers. Ann. Oper. Res. 93 (1–4), 145–176.

Heidemann, D., 1994. Queue length and delay distributions at traffic signals. Trans. Res. Part B 28 (5), 377–389.

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

Heidemann, D., 2001. A queueing theory model of nonstationary traffic flow. Trans. Sci. 35 (4), 405–412.

Heidemann, D., Wegmann, H., 1997. Queueing at unsignalized intersections. Trans. Res. Part B 31 (3), 239–263.

Hillier, F.S., Boling, R.W., 1967. Finite queues in series with exponential or Erlang service times-a numerical approach. Oper. Res. 15 (2), 286–303.

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

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

Jain, R., Smith, J.M., 1997. Modeling vehicular traffic flow using M/G/C/C state dependent queueing models. Trans. Sci. 31 (4), 324–336.

Jun, K.P., Perros, H.G., 1989. Approximate analysis of arbitrary configurations of queueing networks with blocking and deadlock. In: Perros, H.G., Altiok, T. (Eds.), Queueing Networks with Blocking: Proceedings of the First international workshop. North-Holland, Amsterdam, pp. 259–279.

Jun, K.P., Perros, H.G., 1990. An approximate analysis of open tandem queueing networks with blocking and general service times. Eur. J. Oper. Res. 46 (1), 123–135.

Kerbache, L., Smith, J.M., 1987. The generalized expansion method for open finite queueing networks. Eur. J. Oper. Res. 32 (3), 448–461.

Kerbache, L., Smith, J.M., 1988. Asymptotic behaviour of the expansion method for open finite queueing networks. Comp. Oper. Res. 15 (2), 157–169.

Kerbache, L., Smith, J.M., 2000. Multi-objective routing within large scale facilities using open finite queueing networks. Eur. J. Oper. Res. 121 (1), 105–123.

Koizumi, N., Kuno, E., Smith, T.E., 2005. Modeling patient flows using a queueing network with blocking. Health Care Management Sci. 8 (1), 49–60.

Korporaal, R., Ridder, A., Kloprogge, P., Dekker, R., 2000. An analytic model for capacity planning of prisons in the Netherlands. J. Oper. Res. Soc. 51 (11), 1228–1237.

Larson, R., Odoni, A., 1981. Urban Operations Research. Prentice-Hall.

Lighthill, M., Witham, J., 1955. On kinematic waves II. A theory of traffic flow on long crowded roads. Proc. Royal Soc. A 229, 317–345.

Little, J.D.C., 1961. A proof for the queuing formula: L= λ W. Oper. Res. 9 (3), 383–387.

Little, J.D.C., 2011. Little's law as viewed on its 50th anniversary. Oper. Res. 59 (3), 536–549.

MathworksInc., 2012. Optimization Toolbox Version 6.2. User's Guide Matlab. Natick, MA, USA

Meier, P., 2007. Simulation d'un réseau de files d'attente à capacités finies. Technical Report. ROSO Chair of Operations Research SO, Ecole Polytechnique Fédérale de Lausanne.

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

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

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

Osorio, C., Bierlaire, M., 2009. An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. Eur. J. Oper. Res. 196 (3), 996–1007.

Osorio, C., Bierlaire, M., 2012. A tractable analytical model for large-scale congested protein synthesis networks. Eur. J. Oper. Res. 219 (3), 588–597.

Osorio, C., Bierlaire, M., 2013. A simulation-based optimization framework for urban transportation problems. Oper. Res. 61 (6), 1333–1345.

Osorio, C., Chen, X., Marsico, M., Talas, M., Gao, J., Zhang, S., 2014. Reducing gridlock probabilities via simulation-based signal control. In: Proceedings of the International Symposium on Transport Simulation (ISTS). Under review for journal publication, version available at: http://web.mit.edu/osorioc/www/papers/osoChenNYCDOTOfflineSO.pdf.

Osorio, C., Chong, L., 2015. A computationally efficient simulation-based optimization algorithm for large-scale urban transportation problems. Trans. Sci. 49 (3), 623–636.

Osorio, C., Flötteröd, G., 2015. Capturing dependency among link boundaries in a stochastic network loading model. Trans. Sci. 49 (2), 420–431.

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

Osorio, C., Nanduri, K., 2015. Energy-efficient urban traffic management: a microscopic simulation-based approach. Trans. Sci. 49 (3), 637–651.

Osorio, C., Nanduri, K., 2015b. Urban transportation emissions mitigation: coupling high-resolution vehicular emissions and traffic models for traffic signal optimization. Trans. Res. Part B 81, 520–538.

Osorio, C., Yamani, J., 2016. Analytical and scalable analysis of transient tandem markovian finite capacity queueing networks. Trans. Sci.. Forthcoming. Available at: http://web.mit.edu/osorioc/www/papers/osoYamDynAggDisagg.pdf.

Osorio, C., Zhang, C., Flötteröd, G., 2016. Efficient calibration techniques for large-scale traffic simulators. Submitted. Available at: http://web.mit.edu/osorioc/www/papers/osoZhaFlo16Calib.pdf

Rice, J.A., 1994. Mathematical Statistics and Data Analysis. Duxbury Press, Belmont CA USA.

Richards, P.I., 1956. Shock waves on highways. Oper. Res. 4 (1), 42–51.

Schmidt, L.C., Jackman, J., 2000. Modeling recirculating conveyors with blocking. Eur. J. Oper. Res. 124 (2), 422–436.

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

Schweitzer, P.J., 1984. Aggregation methods for large Markov chains. In: Iazeolla, G., Courtois, P.J., Hordijk, A. (Eds.), Mathematical Computer Performance and Reliability, pp. 275–286. North-Holland, Amsterdam
Schweitzer, P.J., Altiok, T., 1989. Aggregate modelling of tandem queues without intermediate buffers. In: Perros, H.G., Altiok, T. (Eds.), Queueing Networks with Blocking: Proceedings of the First international workshop, pp. 47–72. North-Holland, Amsterdam
Singh, A., Smith, J.M., 1997. Buffer allocation for an integer nonlinear network design problem. Comp. Oper. Res. 24 (5), 453–472.
Song, Y., Takahashi, Y., 1991. Aggregate approximation for tandem queueing systems with production blocking. J. Oper. Res. Soc. Japan 34 (3), 329–353.
Stewart, W.J., 2000. Numerical methods for computing stationary distributions of finite irreducible Markov chains. In: Grassmann, W. (Ed.). chapter 4, Computational Probability. Kluwer Academic Publishers, Boston, USa.
Tahilramani, H., Manjunath, D., Bose, S.K., 1999. Approximate analysis of open network of GE/GE/m/N queues with transfer blocking. MASCOTS 0, 164–172.
Takahashi, Y., 1975. A lumping method for numerical calculations of stationary distributions of Markov chains. Res. Reports Inf. Sci. Series B.
Takahashi, Y., 1985. A new type aggregation method for large Markov chains and its application to queueing networks. In: Proceedings of the International Teletraffic Congress 11. Kyoto, Japan
Takahashi, Y., Miyahara, H., Hasegawa, T., 1980. An approximation method for open restricted queuing networks. Oper. Res. 28 (3), 594–602.
Tanner, J.C., 1962. A theoretical analysis of delays at an uncontrolled intersection. Biometrika 49, 163–170.
Texas Transportation Institute, 2012. 2012 Urban Mobility Report. Technical Report. Texas Transportation Institute (TTI), Texas A&M University System.
Tolio, T., Gershwin, S.B., 1998. Throughput estimation in cyclic queueing networks with blocking. Ann. Oper. Res. 79, 207–229.
Transport for London, 2010. Traffic Modelling Guidelines. Version 3.0. Technical Report. Transport for London (TfL).
TSS, 2011. AIMSUN 6.1 Microsimulator User's Manual. Transport Simulation Systems.
Van Woensel, T., Vandaele, N., 2007. Modelling traffic flows with queueing models: a review. Asia-Pacific J. Oper. Res. 24 (4), 1–27.
VSS, 1992. Norme Suisse SN 640837 Installations de feux de circulation; temps transitoires et temps minimaux. Union des professionnels suisses de la route. VSS, Zurich.
van Vuuren, M., Adan, I.J.B.F., Resing-Sassen, S.A.E., 2005. Performance analysis of multi-server tandem queues with finite buffers and blocking. OR Spectrum 27 (2–3), 315–338.
Yperman, I., 2007. The link transmission model for dynamic network loading. Ph.D. thesis Katolieke Universiteit Leuven.