Transportation Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Analytical and Scalable Analysis of Transient Tandem Markovian Finite Capacity Queueing Networks

Carolina Osorio, Jana Yamani

To cite this article:
Carolina Osorio, Jana Yamani (2017) Analytical and Scalable Analysis of Transient Tandem Markovian Finite Capacity Queueing Networks. Transportation Science 51(3):823-840. https://doi.org/10.1287/trsc.2015.0629

http://pubsonline.informs.org/journal/trsc/

TRANSPORTATION SCIENCE
Vol. 51, No. 3, August 2017, pp. 823–840
ISSN 0041-1655 (print), ISSN 1526-5447 (online)

Analytical and Scalable Analysis of Transient Tandem Markovian Finite Capacity Queueing Networks

Carolina Osorio,[a] **Jana Yamani**[a]

[a] Civil and Environmental Engineering Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139
Contact: osorioc@mit.edu (CO); jhy@mit.edu (JY)

Received: November 5, 2013
Revised: August 5, 2014, November 25, 2014
Accepted: December 29, 2014
Published Online in Articles in Advance: January 12, 2017

https://doi.org/10.1287/trsc.2015.0629

Abstract. This paper proposes an analytical model to approximate the transient aggregate joint queue-length distribution of tandem finite (space) capacity Markovian networks. The methodology combines ideas from transient aggregation-disaggregation techniques as well as transient network decomposition methods. The complexity of the proposed method is linear in the number of queues and is independent of the space capacities of the individual queues. This makes it a suitable approach for the analysis of large-scale networks. The transient joint distributions are validated versus simulation estimates. The model is then used to describe urban traffic dynamics and to address a dynamic traffic signal control problem. The signal plan analysis shows the added value of using joint distributional information, and more generally spatial-temporal between-link dependency information, to enhance urban traffic operations.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/trsc.2015.0629.

Keywords: queueing networks • joint distribution • Markovian networks • urban traffic signal control

1. Introduction

With congestion prevailing in urban areas and limited possibilities for road infrastructure expansion, there is a need to rethink how we operate our transportation systems. Transportation strategies are typically formulated such as to improve first-order performance metrics, e.g., expected travel times. They have the potential to further enhance performance by accounting for higher-order distributional information such as to improve, for instance, network reliability and network robustness. Various transportation agencies have recently identified improved network reliability and/or network robustness as critical goals (Texas Transportation Institute 2012; Transport for London 2010; Department of Transportation 2008). Performance measures that account for network reliability/robustness involve the approximation of higher-order distributional information of the main network, or path, performance measures. There are two main challenges that arise when attempting to analytically approximate the full joint network, or path, distribution.

First, an analytical probabilistic approximation of the spatial-temporal dependencies between links (i.e., roads) is needed. Congested urban networks embed intricate traffic dynamics, hence providing an analytical approximation of the between-link interactions is intricate. Hence, the vast majority of the probabilistic network models are simulation based (for a recent review, see Barceló 2010). In the general field of transportation (air, urban, maritime, etc.), few analytical probabilistic and time-dependent traffic models have been developed (Flötteröd and Osorio 2013; Osorio and Flötteröd 2015; Osorio, Flötteröd, and Bierlaire 2011; Gupta 2011; Heidemann 2001; Peterson, Bertsimas, and Odoni 1995a, b; Odoni and Roth 1983). Recent work has proposed link models (Osorio and Flötteröd 2015; Osorio, Flötteröd, and Bierlaire 2011) based on transient Markovian queueing network theory, which are consistent with the mainstream deterministic traffic flow theory models, such as the Kinematic Wave Model (KWM) (Lighthill and Witham 1955; Richards 1956). Such models provide a detailed description of the within-link traffic dynamics. Nonetheless, their use for the joint and tractable analysis of large-scale networks has yet to be explored.

Second, the dimension of the state space of the joint queue-length distribution is exponential in the number of links. Let the state of link i, denoted N_i, be defined as the number of vehicles on the link. Then the network state space is given by $\times_{i\in\mathcal{L}}\{0,1,\ldots,\ell_i\}$, where $\mathcal{L}$ denotes the set of links and ℓ_i is the space capacity of link i. Hence, the dimension of the state space is $\prod_{i\in\mathcal{L}}(\ell_i+1)$. Given the dimensionality of the joint distribution, providing a tractable approximation suitable for the analysis of large-scale networks is a major challenge.

This paper focuses on this dimensionality challenge. It proposes an analytical, tractable, and scalable technique that approximates the joint time-dependent queue-length distribution of a finite (space) capacity tandem (also called series or linear) topology Markovian network. The dimension of the state space of the

proposed method is linear, instead of exponential, in the number of links and is independent of the space capacities of the individual queues. This makes it a suitable approach for the analysis of large-scale tandem networks.

Hereafter, the term capacity refers to space capacity. In the field of queueing network theory, the vast majority of research has focused on stationary analysis, whereas transient techniques have received less attention. Seminal works in transient analysis of a single finite capacity queue include Morse (1958) and Cohen (1982). For recent reviews of transient analysis, see Kaczynski, Leemis, and Drew (2012) and Griffiths, Leonenko, and Williams (2008). For Markovian finite capacity queueing networks (FCQNs), the transient joint queue-length distribution can be obtained by solving a system of linear first-order ordinary differential equations (ODEs) (described in Section 2.1). Closed-form expressions are limited to a single $M/M/1/\ell$ or a single $M/M/2/\ell$ queue (Morse 1958; Sharma and Gupta 1982; Sharma and Shobha 1988). Exact numerical techniques are the most common approach when analyzing transient networks (for reviews, see Stewart 1994, 2009). Nonetheless, the dimension of the joint distribution remains a major challenge.

To address the issue of dimensionality, the most common approach is to decompose the network into subnetworks and approximate the subnetwork distributions. These methods are known as decomposition techniques. A review of stationary decomposition techniques is given in Osorio and Bierlaire (2009). Stationary decomposition methods have mostly decomposed the network into single queues, as in Osorio and Bierlaire (2009). Stationary methods that decompose the network into overlapping subnetworks of three queues, as is done in this paper, include Brandwajn and Jow (1988) and Schmidt and Jackman (2000). Unlike the method proposed in this paper, the latter two methods consider a stationary analysis.

Most transient decomposition techniques assume infinite capacity queues (e.g., McCalla and Whitt 2002; Whitt 1999; Peterson, Bertsimas, and Odoni 1995a; Odoni and Roth 1983). This is due to the complexity of providing an analytical description of the temporal between-queue dependencies in FCQNs, and even more so in congested FCQNs. Transient decomposition techniques for an FCQN include work in the field of manufacturing, where detailed service processes are used to describe intricate machine characteristics; see Li (2005) for general topology networks and Zhang et al. (2013) for tandem topology networks. A technique for general topology Markovian networks is proposed in Flötteröd and Osorio (2013).

A second family of techniques to address the issue of dimensionality are aggregation-disaggregation techniques. The latter describe the state of the network aggregately (i.e., reduced state space), while ensuring consistency with disaggregate (i.e., high-dimensional) distributions (e.g., Schweitzer 1991). Exact transient, and stationary, aggregation-disaggregation techniques have been proposed (Schweitzer 1984). Nonetheless, such approaches are not sufficiently tractable for large-scale networks. An approximate tractable stationary aggregation-disaggregation method appropriate for the analysis of urban networks is proposed in Osorio and Wang (2017).

This paper considers the transient analysis of networks and combines both techniques mentioned above: transient decomposition techniques and transient aggregation-disaggregation techniques. The decomposition technique decomposes the network into overlapping three-queue subnetworks. For each subnetwork, the state of each queue is described aggregately, and an analytical approximation of the between-queue dynamics is proposed. The combination of these two families of ideas leads to a highly tractable and scalable description of network dynamics. It is this combination that leads to a model complexity that is both linear in the number of links (which is often the case of decomposition methods) and independent of the link space capacities (which is often the case of aggregation-disaggregation methods). Additionally, this paper focuses on the transient analysis of networks, unlike the stationary analysis proposed in Osorio and Wang (2017) or Osorio and Bierlaire (2009).

The recently proposed queueing-theoretic Markovian vehicular traffic models that are consistent with deterministic traffic flow theory (Osorio and Flötteröd 2015; Osorio, Flötteröd, and Bierlaire 2011) show the great potential of queueing theory to complement and extend traditional deterministic traffic flow theory. Consistency with the KWM proves the adequacy of using transient Markovian queueing theory to model uninterrupted vehicular traffic, for all levels of congestion. As differentiable and probabilistic models, they can be used as stand-alone models to address a variety of optimization problems. The model proposed in this paper is not formulated such as to be consistent with traditional deterministic traffic flow theory. As part of ongoing work, the aggregation-disaggregation ideas presented in this paper are being combined with detailed traffic-theoretic dynamic link models such as to derive network models that are both consistent with traditional deterministic traffic theories and suitable for the analysis of large-scale networks.

For interrupted traffic (e.g., at signal controlled intersections), stationary or transient Markovian queueing network models that are highly accurate have not been proposed. Nonetheless, they have been successfully used to design computationally efficient simulation-based optimization (SO) algorithms for interrupted urban traffic (Osorio and Bierlaire 2013; Osorio and

Chong 2015; Osorio and Nanduri 2015; Chong and Osorio 2017; Osorio and Selvam 2017). In these SO algorithms, information from high-resolution, yet computationally inefficient, models of interrupted traffic (e.g., stochastic microscopic traffic simulators) is combined with information from low-resolution, yet efficient, analytical Markovian queueing network models. This combination leads to SO algorithms with an appealing resolution-efficiency trade-off.

The model proposed in this paper is formulated such as to be consistent with queueing network theory, rather than traffic flow theory. It is analytical, differentiable, and computationally efficient. Hence, it can be combined with higher-resolution traffic-theoretic models of interrupted traffic to address a variety of optimization problems, both analytical and simulation based. In this paper, the proposed queueing model is used to address an analytical traffic signal control problem (Section 4). The model identifies signal plans with good performance; this shows its potential to be combined with higher-resolution models of interrupted traffic to address intricate time-dependent optimization problems. Additionally, the results of Section 4 show that the signal plans derived by the proposed transient model outperform the signal plans derived by the stationary model used in past work for simulation-based signal control (Osorio and Bierlaire 2013; Osorio and Chong 2015; Osorio and Nanduri 2015). This indicates the potential of the proposed model to enhance the performance of existing SO frameworks.

Modeling and optimizing the spatial and temporal propagation of urban congestion is a great challenge. In particular, models that can describe between-queue dependencies, and more specifically the occurrence and effects of spillbacks are of interest. Major congested cities, such as New York City (Osorio et al. 2015), are rethinking the way they operate their traffic lights such as to mitigate spillbacks. The proposed approach contributes by providing a probabilistic description of between-queue dependencies.

Section 2 presents the proposed methodology. The method is validated versus a general-purpose discrete-event queueing network simulator (Section 3). It is then used to address an urban traffic signal control problem (Section 4), this illustrates its potential to address various transportation optimization problems. Conclusions are presented in Section 5.

2. Methodology

This section is structured as follows. It presents the general transient aggregation-disaggregation framework (Section 2.1). This framework is formulated for an aggregate description of a single queue (Section 2.2), and generalized for a tandem network of queues (Section 2.3). The main challenge in the analytical analysis of a network of finite capacity queues is the analytical description of between-queue dependencies. This challenge is illustrated with a simple example in Section 2.4. The proposed analytical descriptions of the between-queue dependencies are given in Sections 2.5–2.7. An algorithm that summarizes the proposed method is presented in Section 2.8.

2.1. Transient Aggregation-Disaggregation Framework

This paper builds on the exact aggregation-disaggregation technique for transient Markov chains given in Schweitzer (1984). This section presents the main idea underlying the Schweitzer (1984) framework. Consider a continuous-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 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 transition rate matrix, Q, is then defined by

$$Q_{ij} = \begin{cases} q_{ij}, & \text{if } i \neq j, \\ -\sum_{k \in \Omega \setminus i} q_{ik} & \text{if } i = j. \end{cases} \tag{1}$$

Let N denote the network state (e.g., joint network queue length) and let $p_N(t)$ be the row vector that represents the transient joint state distribution at time instant t. Then, $p_N(t)$ satisfies the (forward) Kolmogorov system of equations (see, for instance, Durrett 1999, Chapter 4.2)

$$\frac{dp_N(t)}{dt} = p_N(t)Q. \tag{2}$$

Assuming valid boundary conditions, there are numerous exact numerical techniques to solve the above system of linear first-order ODEs. For reviews on such numerical methods, see Stewart (1994, 2009). The main challenge in solving (2) remains the dimension of the state space. For instance, for a finite capacity queueing network with m queues each with space capacity ℓ, where N represents the joint queue-length state, the state space is of dimension $M = (\ell + 1)^m$, which is exponential in the number of queues and depends on the space capacities of the individual queues.

To address the dimensionality issue, Schweitzer (1984) proposes to partition the M states into $\bar{M}$ aggregate disjoint states, such that $\bar{M} \ll M$. Let $\bar{\Omega}$ denote the set of aggregate states. Let Ω_a denote the set of disaggregate states within aggregate state a. Let A denote the random variable representing the aggregate network state. The probability of being in aggregate state a at time t is denoted $p_{A=a}(t)$ and defined as

$$p_{A=a}(t) = \sum_{i \in \Omega_a} p_{N=i}(t). \tag{3}$$

Schweitzer (1984) shows that the aggregate distribution satisfies a system of the form

$$\frac{dp_{A=a}(t)}{dt} = p_{A=a}(t)\bar{Q}(t), \tag{4}$$

where $\bar{Q}(t)$ represents the transition rate matrix of the aggregate system. Element (a,b) of $\bar{Q}(t)$ is denoted by $\bar{q}_{ab}(t)$ and is referred to as an aggregate transition rate. Schweitzer derives the following exact closed-form expression for $\bar{Q}(t)$ as a function of disaggregate transition rates and disaggregate state probabilities (Schweitzer 1984, Equation (10.4))

$$\bar{q}_{ab}(t) = \begin{cases} \dfrac{\sum_{j\in\Omega_a}\sum_{i\in\Omega_b} p_{N=j}(t)q_{ji}}{\sum_{j\in\Omega_a} p_{N=j}(t)}, \\ \qquad\qquad \text{if } (b,a)\in\bar{\Omega}^2,\ b\neq a, \\ -\displaystyle\sum_{c\in\bar{\Omega}\setminus a}\bar{q}_{ac}, \quad \text{if } a=b. \end{cases} \tag{5}$$

2.2. Aggregate State Representation

This section defines the aggregate state representation. It then considers the aggregation-disaggregation framework presented in Section 2.1, and applies it to a single finite capacity M/M/-type queue. The exact expression derived is used in subsequent sections to formulate the methodology for a network of queues.

Consider a single M/M/1/ℓ queue. The (disaggregate) state of the queue is described by the number of jobs (e.g., vehicles), N, in the queueing system. The (disaggregate) state space is given by $\Omega = \{0,1,\ldots,\ell\}$, where $\ell\in\mathbb{Z}^+$ is the space capacity. Let $\lambda\geq 0$ and $\mu>0$ denote, respectively, the arrival and service rates.

We aggregate the $\ell+1$ (disaggregate) states into the following three (aggregate) states: the queue is empty, the queue is full, and the queue is neither empty nor full. The aggregate states are described by the random variable A: (i) empty queue: $A=0$, $\Omega_0=\{N=0\}$; (ii) nonempty and nonfull queue: $A=1$, $\Omega_1=\{N\in[1,\ell-1]\}$; and (iii) full queue: $A=2$, $\Omega_2=\{N=\ell\}$.

The choice of these three states is based on between-queue dynamics in urban networks, where there are vehicle transmissions from link j to its downstream link k as long as (i) a vehicle is ready to be sent from the upstream link j (i.e., nonempty upstream link: $A_j>0$) and (ii) there is space in the downstream link k to receive a vehicle (i.e., nonfull downstream link: $A_k<2$). With only three states we can describe the boundary conditions that each queue provides to its upstream and downstream queues. This yields a model complexity that is independent of the space capacity of each queue, making this approach highly tractable for large-scale networks. Additionally, the use of such a low-dimensional aggregate description of the within-link state will facilitate the combination of this model with other more detailed link traffic models that describe the within-link dynamics in more detail yet lack tractability (e.g., Osorio and Flötteröd 2015).

The aggregate transition rate matrix of an M/M/1/ℓ queue is given by

$$\bar{Q}(t) = \begin{pmatrix} -\lambda & \lambda & 0 \\ \bar{\mu}(t) & -(\bar{\mu}(t)+\bar{\lambda}(t)) & \bar{\lambda}(t) \\ 0 & \mu & -\mu \end{pmatrix}, \tag{6}$$

where $\bar{\lambda}(t)$ (respectively, $\bar{\mu}(t)$) is used to denote $\bar{q}_{12}(t)$ (respectively, $\bar{q}_{10}(t)$) and represents the rate at which transitions take place from the aggregate state $A=1$ to the full queue state $A=2$ (respectively, empty queue state $A=0$).

The (disaggregate) transition rate matrix of an M/M/1/ℓ queue is given by

$$q_{ij} = \begin{cases} \lambda, & \text{if } j=i+1 \text{ and } i\in[\![0,\ell-1]\!], \\ \mu, & \text{if } j=i-1 \text{ and } i\in[\![1,\ell]\!], \\ -\displaystyle\sum_{j\in\Omega\setminus i} q_{ij}, & \text{if } \quad i=j. \end{cases} \tag{7}$$

Inserting (7) into (5), and noting that $\Omega_0=\{0\}$ and $\Omega_2=\{\ell\}$, we obtain the following exact expressions for the aggregate transition rates:

$$\begin{cases} \bar{\lambda}(t) = \lambda\dfrac{p_{N=\ell-1}(t)}{p_{A=1}(t)}, & \text{(8a)} \\ \bar{\mu}(t) = \mu\dfrac{p_{N=1}(t)}{p_{A=1}(t)}. & \text{(8b)} \end{cases}$$

System (8) is equivalent to

$$\begin{cases} \bar{\lambda}(t) = \lambda p_{N=\ell-1\,|\,A=1}(t), & \text{(9a)} \\ \bar{\mu}(t) = \mu p_{N=1\,|\,A=1}(t). & \text{(9b)} \end{cases}$$

System (9) indicates that an accurate approximation of $\bar{\lambda}(t)$ and of $\bar{\mu}(t)$ can be derived based on an accurate approximation of the probabilities $p_{N=\ell-1\,|\,A=1}(t)$ and $p_{N=1\,|\,A=1}(t)$. We refer to these probabilities as disaggregation probabilities since they represent the probabilities of being in a disaggregate state of a given aggregate state. System (9) will serve as a building block for the proposed methodology.

2.3. Transient Aggregate Description of a Tandem Network

We consider a discrete-time context and introduce the following notation:

- δ time step length;
- k time interval index for interval $[k\delta,(k+1)\delta]$;
- N_i disaggregate state of queue i;
- A_i aggregate state of queue i;
- $\bar{A}_i$ aggregate joint state of subnetwork i: $\bar{A}_i=(A_i,A_{i+1},A_{i+2})$;
- $p^k_{X_i}(t)$ distribution of X_i at continuous time t within time interval k, $t\in[0,\delta]$;
- $\bar{Q}^k_i$ aggregate transition rate matrix of subnetwork i during time interval k.

Consider a tandem topology network with I queues. Each queue has a finite space capacity $\ell_i \in \mathbb{Z}^+$, independent exponentially distributed service times with parameter μ_i, and external arrivals (i.e., arrivals that come from outside of the network) that follow a Poisson process with rate parameter γ_i.

We decompose the network into $I-2$ overlapping subnetworks with three adjacent queues each, as depicted in Figure 1. A three-queue subnetwork is the smallest subnetwork in which the traffic dynamics of each queue account for the states of both its upstream and downstream queues. Subnetwork i consists of three queues indexed $i, i+1$, and $i+2$. The proposed methodology analyses all subnetworks simultaneously, and yields for each subnetwork i an analytical approximation of its transient joint aggregate distribution. For subnetwork i, the joint aggregate state probabilities at continuous-time t of time interval k are denoted $p^k_{\bar{A}_i=s}(t)$, where an aggregate state s is defined as the triplet: $s = (j_i, j_{i+1}, j_{i+2}) \in \{0,1,2\}^3$. Each queue of a subnetwork has three aggregate states, hence the dimension of the state space of the subnetwork is aggregated into $3^3 = 27$ distinct states. For a network with I queues, the proposed approach yields $I-2$ subnetwork distributions, each with a state space of dimension 27. Hence, the complexity of the proposed model is linear, instead of exponential, in the number of queues and is independent of the space capacities of the individual queues. This makes it a suitable approach for the analysis of large-scale tandem networks.

For each subnetwork i, we assume that the temporal evolution of its joint aggregate distribution satisfies a system of equations of the form (4). Additionally, for a given time interval k of duration δ, we approximate the aggregate transition rate matrix of subnetwork i, $\bar{Q}_i(t)$, by a time invariant matrix $\bar{Q}^k_i$. Equation (4) then becomes a linear ODE

$$\frac{dp^k_{\bar{A}_i}(t)}{dt} = p^k_{\bar{A}_i}(t)\bar{Q}^k_i, \quad \forall\, t \in [0,\delta], \tag{10}$$

which has a solution of the form (see, for instance, Reibman 1991)

$$p^k_{\bar{A}_i}(t) = p^k_{\bar{A}_i}(0)e^{t\bar{Q}^k_i}, \quad \forall\, t \in [0,\delta]. \tag{11}$$

The initial conditions that ensure the temporal continuity of the aggregate distribution across time intervals are given by

$$p^k_{\bar{A}_i}(0) = p^{k-1}_{\bar{A}_i}(\delta). \tag{12}$$

The approximation of the aggregate time-dependent transition rate matrix $\bar{Q}_i(t)$, is formulated as a function $f_{\bar{Q}}$ of four parameters, three of which are time dependent

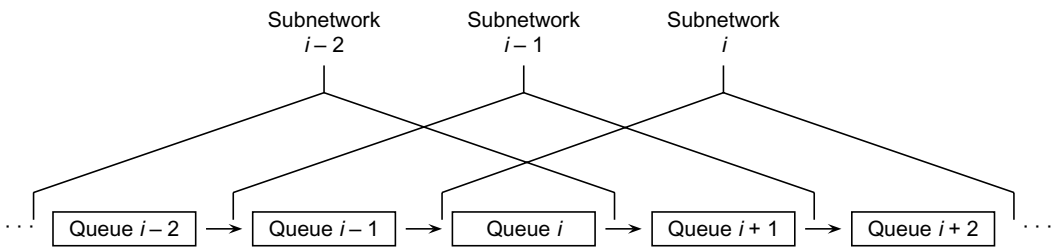
$$\bar{Q}^k_i = f_{\bar{Q}}(\bar{\gamma}^k_i, \bar{\mu}^k_i, \alpha^k_i, \beta_i), \tag{13}$$

where $\bar{\gamma}^k_i$ represents the rates of arrival from outside the subnetwork, $\bar{\mu}^k_i$ denotes subnetwork service rates, α^k_i are disaggregation probabilities, and β_i are blocking probabilities. The full expression for $\bar{Q}^k_i$ (i.e., the definition of the function $f_{\bar{Q}}$) is given in Table 1 of Online Appendix A. The structure of the matrix $\bar{Q}^k_i$ is the same as that of the time-independent transition rate matrix used in Osorio and Wang (2017, Table 10). The definitions and approximations of α^k_i, β_i, $\bar{\gamma}^k_i$, and $\bar{\mu}^k_i$ are described, respectively, in Sections 2.5–2.7. Section 2.4 illustrates through an example the intricate traffic phenomena that may arise in finite capacity networks. This serves to highlight the challenge of approximating these subnetwork parameters.

2.4. Describing the Propagation of Congestion Through Blocking

When considering a network of multiple finite capacity queues, intricate traffic dynamics may arise because of the emersion of blocking (referred to as spillback in urban traffic). Blocking arises when a job (e.g., a vehicle) completes service yet finds no available space in its downstream queue to proceed. Hence, the job is said to be blocked by its downstream queue. A blocked job is also blocking the use of the underlying server (e.g., road space) by other upstream jobs. There are various types of blocking mechanisms (see Balsamo, De Nitto Persone, and Onvural 2001), here we consider blocking

Figure 1. Overlapping Subnetworks of Three Tandem Queues



after service, which is also known as production blocking or manufacturing blocking. In this case, once a job is blocked it continues to occupy the underlying server until it can proceed downstream (i.e., until it is unblocked). This form of blocking mimics well the spillback dynamics that arise in urban traffic.

Blocking leads to intricate between-queue dependencies. For instance, 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. Additionally, for a general topology network if queue i is blocked by downstream queue j, then queue j is full and may be blocking jobs at other upstream queues other than queue i. Hence, the rate of job departures from queue i (known as the unblocking rate) depends not only on the state and service rate of queue j but also on the occurrence of blocking at all upstream queues of queue j.

The following example, taken from Osorio and Wang (2017, Section 2.3.2), illustrates the notion of blocking and the intricate between-queue dependencies that it leads to. Consider for subnetwork i a joint aggregate state $s = (1,2,2)$, where queue i (i.e., the most upstream queue) is in state 1, and queues $i+1$ and $i+2$ are in 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 is either $(1,2,2)$ (this occurs with probability $p_{N_i>1|A_i=1}$) or $(0,2,2)$ (this occurs with probability $p_{N_i=1|A_i=1}$). These probabilities are known as disaggregation probabilities.

This example illustrates the need to approximate: (i) disaggregation probabilities, and (ii) blocking probabilities for states where blocking can occur. Analytical approximations for these two elements are proposed, respectively, in Sections 2.5 and 2.6. One of the main challenges when analyzing finite capacity networks is to accurately approximate blocking and unblocking events. This is an even greater challenge in our context, since the proposed paper considers an aggregate (i.e., nondetailed) representation of queue states.

2.5. Disaggregation Probabilities

For a three-queue network, an exact expression for the aggregate and disaggregate transition rates can be derived as was done for a single queue in Section 2.1 (which lead to System (9)). The aggregate transition rate matrix is then described as a function of disaggregation probabilities (see System (9)), where each queue j in subnetwork i has two disaggregation probabilities that are of interest: $p_{N_j=n|A_j=1}(t)$, $n \in \{1, \ell_j - 1\}$.

We propose to approximate these disaggregation probabilities by accounting for the joint subnetwork state. In other words, we approximate $p_{N_j=n|A_j=1}(t)$ by using information from $p_{N_j=n|A_j=1,\bar{A}_i=s}(t)$. That is, we derive state-dependent disaggregation probabilities. Let us describe this in more detail.

For subnetwork i, we consider a total of six scenarios (or sets of states) described below. These scenarios consider each queue of the subnetwork and distinguish between states where the queue can be blocked and if so by which queue.

For queue i (which is the most upstream queue in subnetwork i), we consider three types of disaggregation probabilities:

(1) If its directly downstream queue $i+1$ is not full, then queue i cannot be blocked. This leads to the following disaggregation probabilities:

$$p_{N_i=n|A_i=1,A_{i+1}\neq 2}(t), \quad n \in \{1, \ell_i - 1\}.$$

(2) If queue $i+1$ is full but queue $i+2$ is not full, then queue i can only be blocked by queue $i+1$:

$$p_{N_i=n|A_i=1,A_{i+1}=2,A_{i+2}\neq 2}(t), \quad n \in \{1, \ell_i - 1\}.$$

(3) If both queues $i+1$ and $i+2$ are full, then queue i can be blocked by either queue:

$$p_{N_i=n|A_i=1,A_{i+1}=2,A_{i+2}=2}(t), \quad n \in \{1, \ell_i - 1\}.$$

Similarly for queue $i+1$:

(4) If its downstream queue $i+2$ is not full, then queue $i+1$ cannot be blocked:

$$p_{N_{i+1}=n|A_{i+1}=1,A_{i+2}\neq 2}(t), \quad n \in \{1, \ell_{i+1} - 1\}.$$

(5) If its downstream queue $i+2$ is full, then queue $i+1$ can be blocked by queue $i+2$:

$$p_{N_{i+1}=n|A_{i+1}=1,A_{i+2}=2}(t), \quad n \in \{1, \ell_{i+1} - 1\}.$$

For the most downstream queue of subnetwork i, queue $i+2$, we consider a single case:

(6) Queue $i+2$ cannot be blocked:

$$p_{N_{i+2}=n|A_{i+2}=1}(t), \quad n \in \{1, \ell_{i+2} - 1\}.$$

The above description presents the six scenarios that we consider. For each scenario, we propose an approximation for the corresponding disaggregation probabilities.

Note from the above description of six scenarios that for subnetwork i the most detailed description of blocking is given for queue i. This is because its blocking scenarios account for joint states with two of its downstream queues (queues $i+1$ and $i+2$), whereas for queue $i+1$ the state of only one downstream queue is

accounted for, and for queue $i+2$ no information from its downstream queues are accounted for. Thus, we propose an approach where the disaggregation probabilities of a given queue i are derived by analyzing subnetwork i (i.e., the subnetwork where queue i is the most upstream queue). In other words, for subnetwork i the disaggregation probabilities corresponding to queue i (i.e., scenarios 1, 2, and 3) are obtained from the analysis of subnetwork i. This is described in Section 2.5.1. For subnetwork i, the disaggregation probabilities of queues $i+1$ and $i+2$ are obtained from the analysis of subnetworks $i+1$ and $i+2$, as described in Section 2.5.2.

2.5.1. Scenarios 1–3. For subnetwork i, the disaggregation probabilities of queue i correspond to scenarios $j \in \{1,2,3\}$. Let us describe how these disaggregation probabilities are approximated. They each have the form $p_{N_i=n\,|\,E_j}(t)$, $n \in \{1, \ell_i - 1\}$, where E_j denotes the conditioning event of scenario j. Considering a discrete time context, we approximate each of these probabilities by a constant value during time interval k, denoted $\alpha^k_{i,j,n}$ and approximated by

$$\alpha^k_{i,j,n} = p^{k-1}_{N_i=n\,|\,E_j}(\delta), \quad j \in \{1,2,3\},\ n \in \{1, \ell_i - 1\}. \tag{14}$$

Recall from Section 2.3 that our method approximates the aggregate subnetwork distributions $p_{\bar{A}_i}$. Hence at the beginning of time interval k the aggregate joint distribution $p^{k-1}_{\bar{A}_i\,|\,E_j}$ and the aggregate marginal distributions $p^{k-1}_{A_i\,|\,E_j}$ are known, but the disaggregate distribution that appears in the right-hand side of (14), $p^{k-1}_{N_i\,|\,E_j}$, is unknown.

To approximate this unknown distribution, we assume it has the same functional form as that of the disaggregate queue-length distribution of a single isolated M/M/1/ℓ queue. The functional form of the disaggregate distribution for a single queue is derived in Morse (1958, pp. 65–67). Its expression for a given queue with space capacity ℓ, arrival rate λ, service rate μ, and initial distribution $p_N(0)$, is given by $\forall\, n = 0,1,\ldots,\ell, \forall\, t \in [0,\delta]$

$$\begin{cases} p_{N=n}(t) = \displaystyle\sum_{m=0}^{\ell} p_{N=m}(0) d^m_n(t,\lambda,\mu,\ell); & \text{(15a)} \\ d^m_n(t,\lambda,\mu,\ell) = s_n + \dfrac{2\rho^{(n-m)/2}}{\ell+1} \displaystyle\sum_{j=1}^{\ell} \frac{\mu}{x_j} \\ \qquad \cdot \left[\sin\dfrac{jm\pi}{\ell+1} - \sqrt{\rho}\sin\dfrac{j(m+1)\pi}{\ell+1}\right] \cdots \\ \qquad \cdot \left[\sin\dfrac{jn\pi}{\ell+1} - \sqrt{\rho}\sin\dfrac{j(n+1)\pi}{\ell+1}\right] e^{-x_j t}; & \text{(15b)} \\ s_n = \dfrac{1-\rho}{1-\rho^{\ell+1}}\rho^n; & \text{(15c)} \\ x_j = \lambda + \mu - 2\sqrt{\lambda\mu}\cos\dfrac{j\pi}{\ell+1}; & \text{(15d)} \\ \rho = \lambda/\mu. & \text{(15e)} \end{cases}$$

We denote the above system of equations as a function f_D

$$p_{N=n}(t) = f_D(n,t,\lambda,\mu,\ell,p_N(0)). \tag{16}$$

The distribution $p^{k-1}_{N_i\,|\,E_j}(t)$ (i.e., $\{p^{k-1}_{N_i=n\,|\,E_j}(t), n \in \{0,1,\ldots,\ell_i\}\}$) is approximated by assuming it satisfies (15), i.e.,

$$p^{k-1}_{N_i=n\,|\,E_j}(t) = f_D\big(n,t,\lambda^{k-1}_{i,j},\mu^{k-1}_{i,j},p^{k-2}_{N_i\,|\,E_j}(\delta)\big). \tag{17}$$

In (17) the parameters $\lambda^{k-1}_{i,j}$ and $\mu^{k-1}_{i,j}$ are unknown. They are approximated by noticing that there is a one-to-one mapping between the disaggregate state $N_i = 0$ (respectively, $N_i = \ell_i$) and the aggregate state $A_i = 0$ (respectively, $A_i = 2$). Ensuring consistency among the disaggregate and the aggregate probabilities of these states leads to the following equations:

$$\begin{cases} p^{k-1}_{A_i=0\,|\,E_j}(\delta) = p^{k-1}_{N_i=0\,|\,E_j}(\delta), & \text{(18a)} \\ p^{k-1}_{A_i=2\,|\,E_j}(\delta) = p^{k-1}_{N_i=\ell_i\,|\,E_j}(\delta). & \text{(18b)} \end{cases}$$

Thus, we can obtain the parameters $\lambda^{k-1}_{i,j}$ and $\mu^{k-1}_{i,j}$ by solving the following system of equations:

$$\begin{cases} p^{k-1}_{A_i=0\,|\,E_j}(\delta) = f_D(0,\delta,\lambda^{k-1}_{i,j},\mu^{k-1}_{i,j},p^{k-2}_{N_i\,|\,E_j}(\delta)), & \text{(19a)} \\ p^{k-1}_{A_i=2\,|\,E_j}(\delta) = f_D(\ell_i,\delta,\lambda^{k-1}_{i,j},\mu^{k-1}_{i,j},p^{k-2}_{N_i\,|\,E_j}(\delta)). & \text{(19b)} \end{cases}$$

Let us detail this. Recall that f_D represents the system of equations (15). In the system of equations (19) the fixed input parameters are 0, ℓ, δ, and $p^{k-2}_{N_i\,|\,E_j}$; there are two endogenous variables (i.e., the unknowns in the system of equations): $\lambda^{k-1}_{i,j}$, $\mu^{k-1}_{i,j}$. In other words, (19) represents a two-dimensional system of nonlinear equations.

Given the rates $\lambda^{k-1}_{i,j}$ and $\mu^{k-1}_{i,j}$, the distribution $p^{k-1}_{N_i\,|\,E_j}(t)$ is fully defined, and is used to evaluate the disaggregation probabilities: $p^{k-1}_{N_i=n\,|\,E_j}(t)$, $n \in \{1, \ell_i - 1\}$, i.e., $\alpha^k_{i,j,n}$, $j \in \{1,2,3\}$, $n \in \{1, \ell_i - 1\}$.

2.5.2. Scenarios 4–6. Section 2.5.1 describes the method to obtain for all subnetworks i the probabilities $\alpha^k_{i,j,n}$, $j \in \{1,2,3\}$. This section describes the approximation of the remaining disaggregation probabilities, i.e., $\alpha^k_{i,j,n}$, $j \in \{4,5,6\}$. Our proposed network decomposition consists of overlapping subnetworks. Hence, a queue may belong to multiple subnetworks. For instance, queue i belongs to subnetworks $i-2$, $i-1$, and i. The remaining disaggregation probabilities (i.e., $\alpha^k_{i,j,n}$, $j \in \{4,5,6\}$) are

derived such as to ensure consistency among the disaggregation probabilities of a given queue i across subnetworks. The following equations ensure consistency:

$$\begin{cases} \alpha^k_{i,4,n} = \alpha^k_{i+1,1,n}, & n \in \{1, \ell_{i+1} - 1\}; \quad (20a) \\ \alpha^k_{i,5,n} = p^{k-1}_{A_{i+3}\neq 2}(\delta)\alpha^k_{i+1,2,n} + p^{k-1}_{A_{i+3}=2}(\delta)\alpha^k_{i+1,3,n}, & n \in \{1, \ell_{i+1} - 1\}; \quad (20b) \\ \alpha^k_{i,6,n} = p^{k-1}_{A_{i+3}\neq 2}(\delta)\alpha^k_{i+2,1,n} + p^{k-1}_{A_{i+3}=2}(\delta) \cdot [p^{k-1}_{A_{i+4}\neq 2}(\delta)\alpha^k_{i+2,2,n} + p^{k-1}_{A_{i+4}=2}(\delta)\alpha^k_{i+2,3,n}], \ldots & n \in \{1, \ell_{i+2} - 1\}. \quad (20c) \end{cases}$$

The left-hand side of Equation (20a) considers scenario 4 of subnetwork i. That scenario considers queue $i+1$ and assumes that its directly downstream queue $(i+2)$ is not full. This is equivalent to considering scenario 1 of subnetwork $i+1$, which is the left-hand side of Equation (20a). Similarly, Equation (20b) is derived. Equation (20c) is obtained by defining $\alpha^k_{i,6,n}$ just as $\alpha^k_{i,5,n}$ in (20b)

$$\alpha^k_{i,6,n} = p^{k-1}_{A_{i+3}\neq 2}(\delta)\alpha^k_{i+1,4,n} + p^{k-1}_{A_{i+3}=2}(\delta)\alpha^k_{i+1,5,n}, \quad (21)$$

and then inserting the expressions of $\alpha^k_{i+1,4,n}$ (respectively, $\alpha^k_{i+1,5,n}$) as given by (20a) (respectively, (20b)). In System (20), the marginal probabilities of a given queue i, $p_{A_i}(\delta)$, are derived from the analysis of network $i-2$.

2.6. Blocking Probabilities

Considering the set of states where jobs can be blocked, we approximate the corresponding blocking probabilities with state-dependent, simple, and exogenous expressions. These are given in Table 1. These expressions are taken from Osorio and Wang (2017, Section 2.3.2). This table considers the queues of subnetwork i 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). Multiple states for the initial joint states are listed in braces. For instance, the first row considers the case where queue i can be blocked by queue $i+1$ and cannot be blocked by queue $i+2$. This can occur as long as queue i is nonempty ($A_i \in \{1,2\}$), queue $i+1$ is full ($A_{i+1} = 2$), and queue $i+2$ is not full ($A_{i+2} \in \{0,1\}$). The approximation of all blocking probabilities (column 4) are given by simple expressions that involve only the exogenous parameters μ_i, $i = 1, \ldots, I$. The approximation is based on the property referred to as "competing exponentials" or "competing Poisson processes." Consider n independent exponentially distributed random variables $\{X_r\}_{r=1:n}$ with rate parameter μ_r, then

$$P(X_r < X_i, \forall\, i \neq r) = \frac{\mu_r}{\sum_{j=1}^n \mu_j}. \quad (22)$$

For a derivation, see Larson and Odoni (1981, Chapter 2.12.4, Equation (2.62)). Hence, if we consider n independent services, the probability that the first service completion is of type r is given by Equation (22). This property is used to approximate the blocking probabilities in column 4 of Table 1. For instance, the first row of the table considers states where queue i can be blocked by queue $i+1$ and not by queue $i+2$. This can occur if queue i is nonempty, queue $i+1$ is full, queue $i+2$ is not full, and queue i finishes service before queue $i+1$. The probability that queue i finishes service before queue $i+1$ is $\mu_i/(\mu_i + \mu_{i+1})$.

2.7. Subnetwork Arrival and Service Rates

Subnetwork i is a part of a larger network, hence the arrival rate to its most upstream queue (queue i) depends on the states and rates of queues further upstream of the subnetwork (e.g., queue $i-1$). The total external arrival rate (i.e., from outside the subnetwork) to the queues of subnetwork i (during time interval k) is denoted $\bar{\gamma}^k_i$ and is given by

$$\bar{\gamma}^k_i = [\hat{\gamma}^k_i, \gamma_{i+1}, \gamma_{i+2}], \quad (23)$$

where $\bar{\gamma}^k_i$ is a three-dimensional vector and each term in the brackets is a scalar. The rates γ_{i+1} and γ_{i+2} are exogenous parameters. The rate $\hat{\gamma}^k_i$ is approximated by

$$\hat{\gamma}^k_i(1 - p^{k-1}_{A_i=2}(\delta)) = \gamma_i + \hat{\gamma}^k_{i-1}(1 - p^{k-1}_{A_{i-1}=2}(\delta)). \quad (24)$$

The above expression is a flow conservation equation that relates the arrival rate to queue i, $\hat{\gamma}^k_i$, to its external

Table 1. Blocking Probabilities of Subnetwork i

Blocked queues	Source queue	Initial joint states $\bar{A}_i$	Blocking probability
i	$i+1$	$(\{1,2\},2,\{0,1\})$	$\beta_{i,1} = \dfrac{\mu_i}{\mu_i + \mu_{i+1}}$
$i, i+1$	$i+2$	$(\{1,2\},2,2)$	$\beta_{i,2} = \dfrac{\mu_i}{\mu_i + \mu_{i+1} + \mu_{i+2}} \dfrac{\mu_{i+1}}{\mu_{i+1} + \mu_{i+2}} + \dfrac{\mu_{i+1}}{\mu_i + \mu_{i+1} + \mu_{i+2}} \dfrac{\mu_i}{\mu_i + \mu_{i+2}}$
$i+1$	$i+2$	$(\{0,1,2\},1,2), (0,2,2)$	$\beta_{i,3} = \dfrac{\mu_{i+1}}{\mu_{i+1} + \mu_{i+2}}$
$i+1$	$i+2$	$(\{1,2\},2,2)$	$\beta_{i,4} = \dfrac{\mu_{i+1}}{\mu_i + \mu_{i+1} + \mu_{i+2}} \dfrac{\mu_{i+2}}{\mu_i + \mu_{i+2}}$

arrival rate (from outside the network), γ_i, and to the arrival rate of its upstream queue, $\hat{\gamma}_{i-1}$. The probabilities arise because we consider finite space capacity models ($\ell < +\infty$). For such models, flow can enter the queue as long as it is not full, hence the flow that enters is the product of the total arrival rate, $\hat{\gamma}_i^k$, with the probability of the queue not being full, $1 - p_{A_i=2}^{k-1}$. Equation (24) is a time-dependent extension of the time-independent subnetwork arrival rate proposed in Osorio and Wang (2017, Equation (20)).

Similarly, since subnetwork i is a part of a larger network, the service rate of its most downstream queue (queue $i+2$) depends on the states and rates of queues further downstream of the subnetwork (e.g., queue $i+3$). When analyzing subnetwork i the service rate vector of its queues is denoted $\bar{\mu}_i$ and is given by

$$\bar{\mu}_i^k = [\mu_i, \mu_{i+1}, \hat{\mu}_{i+2}^k], \tag{25}$$

where $\bar{\mu}_i^k$ is a three-dimensional vector and each term in the brackets is a scalar. The rates μ_i and μ_{i+1} are exogenous parameters. The rate $\hat{\mu}_i^k$ is approximated by

$$\frac{1}{\hat{\mu}_i^k} = \frac{1}{\mu_i} + \left[p_{A_{i+1}=2}^{k-1}(\delta) \frac{\mu_i}{\mu_i + \mu_{i+1}} \right] \cdot \left[\frac{\hat{\gamma}_{i+1}^k (1 - p_{A_{i+1}=2}^{k-1}(\delta))}{\hat{\gamma}_i^k (1 - p_{A_i=2}^{k-1}(\delta))} \frac{1}{\hat{\mu}_{i+1}^k} \right]. \tag{26}$$

This expression relates the effective service rate of queue i, $\hat{\mu}_i^k$, to its exogenous service rate, μ_i, plus a term that approximates the expected blocking time. The expression in the first pair of brackets represents the probability that a job (e.g., a vehicle) in queue i gets blocked. This is approximated by the product of (i) the probability that the downstream queue is full $p_{A_{i+1}=2}$, and (ii) the probability that the service at queue i is completed before the service of the downstream queue $i+1$. The expression in the second pair of brackets represents the expected blocked time of a job at queue i given that it gets blocked. The left fraction represents the inverse of the proportion of arrivals to the downstream queue that arise from queue i (this may not be equal to 1 since external arrivals from outside the network are allowed). The right fraction represents the expected time between unblocking events, which is given by the inverse of the effective service rate of the downstream queue $\hat{\mu}_{i+1}$. Equation (26) is a time-dependent extension of the time-independent expression proposed in Osorio and Wang (2017, Equations (30), (32), and (33)).

2.8. Algorithm

Algorithm 1 summarizes the proposed approach. The algorithm involves solving three systems of equations at steps 6(a), 6(e), and 6(g), respectively. The system of step 6(e) is a system that is linear in the unknowns $\hat{\gamma}^k$. The system of step 6(g) is linear in the unknowns $1/\hat{\mu}^k$. Step 6(a) solves a set of two-dimensional nonlinear system of equations. These are solved with the MATLAB routine *fsolve* and its "trust region reflective" algorithm (Coleman and Li 1994, 1996). The termination tolerance on the function value is set to 10^{-6}.

Algorithm 1 (Tandem network algorithm).

Carry out each of the following steps for all subnetworks i before proceeding to the next step.

1. Set the exogenous parameters μ, γ, ℓ.
2. Evaluate the exogenous blocking probabilities β_i according to Table 1.
3. Set $k = 1$.
4. Set initial aggregate joint distributions $p_{\bar{A}_i}^k(0)$ (or equivalently $p_{\bar{A}_i}^{k-1}(\delta)$).
5. Set initial conditional disaggregate distributions: $p_{N_i|E_j}^{k-1}(\delta)$. Go to step 6(c).
6. Repeat the following for time intervals $k = 1, 2, \ldots$.
   (a) Compute $\lambda_{i,j}^{k-1}$ and $\mu_{i,j}^{k-1}$, $j \in \{1,2,3\}$ by solving the system of equations (19).
   (b) Compute the disaggregate distributions $p_{N_i|E_j}^{k-1}$, $j \in \{1,2,3\}$ according to (17) and (15).
   (c) Compute the disaggregation probabilities $\alpha_{i,j,n}^k$, $j \in \{1,2,3\}$, $n \in \{1, \ell_i - 1\}$ according to (14).
   (d) Evaluate $\alpha_{i,j,n}^k$, $j \in \{4,5,6\}$, $n \in \{1, \ell_i - 1\}$ according to (20).
   (e) Solve the linear system of equations (24) to obtain $\hat{\gamma}^k$.
   (f) Compute $\bar{\gamma}_i^k$ according to (23).
   (g) Solve the system of equations (26) to obtain $\hat{\mu}^k$.
   (h) Compute $\bar{\mu}_i^k$ according to (25).
   (i) Evaluate the aggregate transition rate matrix $\bar{Q}_i^k$ according to (13), where the function $f_{\bar{Q}}$ is given by Table 1 of Online Appendix A.
   (j) Evaluate the aggregate joint distribution at the end of the time interval $p_{\bar{A}_i}^k(\delta)$ according to (11).
   (k) Set initial conditions for the next time interval: $p_{\bar{A}_i}^{k+1}(0) = p_{\bar{A}_i}^k(\delta)$.

3. Validation

We validate the transient aggregate joint distributions versus distributions estimated with a discrete event simulator of a Markovian FCQN (Meier 2007). For more extensive validation experiments and details, we refer the reader to Yamani (2013, Chapter 3).

The simulated estimates are obtained from 10,000 simulation replications. Let $p_s(t)$ denote the transient probability of being in a given joint aggregate state s at time t. A 95% confidence interval for $p_s(t)$ is given by $\hat{p}_s(t) \pm 1.96\sqrt{\hat{p}_s(t)(1 - \hat{p}_s(t))/(10{,}000 - 1)}$, where $\hat{p}_s(t)$ is the simulated estimate of $p_s(t)$ (see, for instance, Rice 1994, Section 7.3.3). We collect simulated estimates

Table 2. Validation Scenarios for Three Queue Network

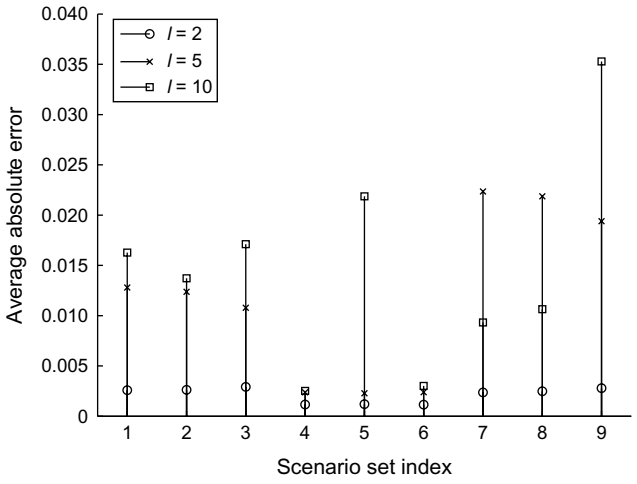
Scenario	$[\mu_1, \mu_2, \mu_3]$	$[\ell_1, \ell_2, \ell_3]$
1	[1.9,1.9,1.9]	[2,2,2]
2	[1.9,1.9,1.9]	[5,5,5]
3	[1.9,1.9,1.9]	[10,10,10]
4	[1.9,4,6]	[2,2,2]
5	[1.9,4,6]	[5,5,5]
6	[1.9,4,6]	[10,10,10]
7	[6,4,1.9]	[2,2,2]
8	[6,4,1.9]	[5,5,5]
9	[6,4,1.9]	[10,10,10]
10	[1.7,1.7,1.7]	[2,2,2]
11	[1.7,1.7,1.7]	[5,5,5]
12	[1.7,1.7,1.7]	[10,10,10]
13	[1.7,4,6]	[2,2,2]
14	[1.7,4,6]	[5,5,5]
15	[1.7,4,6]	[10,10,10]
16	[6,4,1.7]	[2,2,2]
17	[6,4,1.7]	[5,5,5]
18	[6,4,1.7]	[10,10,10]
19	[2,2,2]	[2,2,2]
20	[2,2,2]	[5,5,5]
21	[2,2,2]	[10,10,10]
22	[2,4,6]	[2,2,2]
23	[2,4,6]	[5,5,5]
24	[2,4,6]	[10,10,10]
25	[6,4,2]	[2,2,2]
26	[6,4,2]	[5,5,5]
27	[6,4,2]	[10,10,10]

with a time step of $t = 1$. The analytical model is run with time step $\delta = 0.1$. For all validation scenarios, we consider an initially empty network. In most of these scenarios stationarity is reached by time $t = 50$. Stationarity is assumed to be reached if the Euclidean distance between the simulated distributions across two consecutive intervals is below 10^{-7}.

We consider a tandem topology network with three queues. External arrivals arise only to the first (i.e., most upstream) queue, with $\gamma_1 = 1.8$. We consider a set of 27 scenarios tabulated in Table 2. All scenarios consider highly congested traffic conditions. Across the scenarios we vary the minimal service rate μ_i, leading to a maximal ratio γ_1/μ_i that takes values $\{0.9, 0.95, 1.05\}$. We also vary the location of the queue with the highest traffic intensity (we call this the bottleneck queue): it can be either the first queue (most upstream), the last queue (most downstream), or all three queues. For a given scenario, all three queues have common space capacity, ℓ_i. Across the scenarios, the space capacity ℓ_i can take values $\{2, 5, 10\}$. We consider all combinations of the three possible locations of the bottleneck, the three values of the bottleneck traffic intensity, and the three space capacity values. This leads to a total of $3^3 = 27$ scenarios.

Figure 2 calculates for each of the 27 scenarios a single error metric, which is the average absolute error.

Figure 2. Average Absolute Error for Each of the 27 Scenarios, Which Are Grouped Such as to Observe the Effect of Varying the Space Capacity ℓ



The average is taken over all aggregate state probabilities of all queues at all time steps ($t = 1, 2, \ldots, 50$). Each average is an average over a total of 1,350 state probabilities. The total 1,350=$3^3 \cdot 50$ corresponds to the 27 joint aggregate state probabilities of the three-queue network, evaluated for each of the 50 time instances.

In Figure 2 the circles (respectively, crosses and squares) denote the scenarios where the queues have a space capacity $\ell = 2$ (respectively, $\ell = 5$ and $\ell = 10$). Figure 2 groups the 27 scenarios of Table 3 of Online Appendix C into nine sets (indexed 1 to 9 along the x-axis of Figure 2). For a given scenario set (i.e., a given x-value in Figure 2), the only difference in the three scenarios is their space capacity value, all other scenario parameters are common.

The first three sets of scenarios (indexed 1, 2, and 3 in Figure 2) correspond to the cases where all queues have common traffic intensities. The index increases as the traffic intensity increases; i.e., index 1 (respectively, 2 and 3) corresponds to a traffic intensity of 0.9 (respectively, 0.95 and 1.05) for all queues. The second three sets of scenarios (indexed 4, 5, and 6) correspond to the cases where the bottleneck queue (i.e., the queue with the highest traffic intensity) is the most upstream queue. Again, the index increases as the traffic intensity of the bottleneck queue increases; i.e., index 4 (respectively, 5 and 6) correspond to a traffic intensity of the bottleneck queue of 0.9 (respectively, 0.95 and 1.05). The final three sets of scenarios (indexed 7, 8, and 9) correspond to the cases where the bottleneck queue is the most downstream queue. Again, the index increases as the traffic intensity of the bottleneck queue increases; i.e., index 7 (respectively, 8 and 9) corresponds to a traffic intensity of the bottleneck queue of 0.9 (respectively, 0.95 and 1.05).

For seven out of the nine sets of scenarios (i.e., all sets except 7 and 8) the average absolute error increases

Figure 3. Average Absolute Error for Each of the 27 Scenarios, Which Are Grouped Such as to Observe the Effect of Varying the Traffic Intensity of the Bottleneck Queue

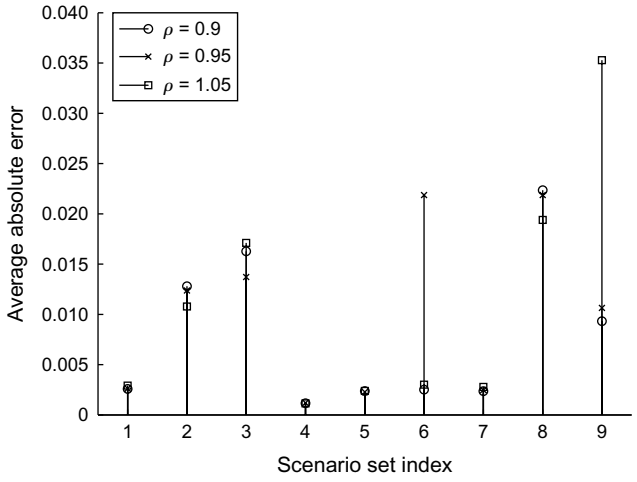
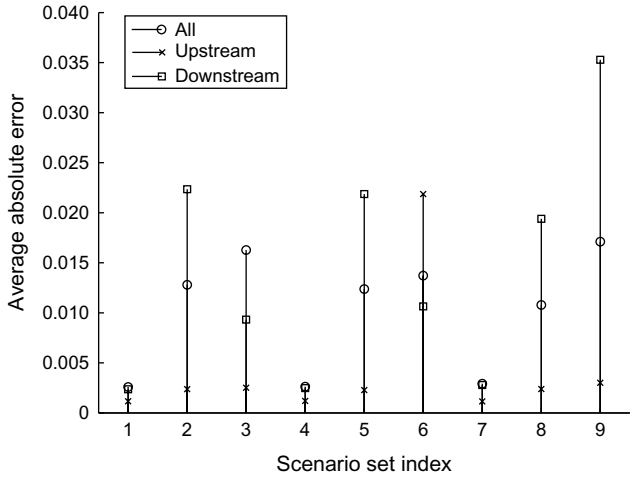


Figure 4. Average Absolute Error for Each of the 27 Scenarios, Which Are Grouped Such as to Observe the Effect of Varying the Location of the Bottleneck Queue



with the space capacity. This can be seen in Figure 2 as follows: for a given scenario set, the lowest average corresponds to the circle ($\ell = 2$), followed by the cross ($\ell = 5$), and then followed by the square ($\ell = 10$). Figure 2 shows that the sets of scenarios with the smallest errors are sets 4 and 6, which both correspond to cases where the bottleneck location is upstream. This is further illustrated in the remaining figures of this section.

Figure 3 also considers for each scenario the average absolute error; it groups the scenarios according to common values of the highest traffic intensity. For a given scenario set (i.e., a given x-value in Figure 3), the only difference in the three scenarios is the value of the bottleneck traffic intensity value; all other scenario parameters are common. The circles (respectively, crosses and squares) denote the scenarios where the bottleneck queue has a traffic intensity of 0.9 (respectively, 0.95 and 1.05). For seven of the nine sets of scenarios, the error does not vary much with the traffic intensity. This holds for all sets except sets 6 and 9. Note that sets 6 and 9 both consider scenarios where the queues have the largest space capacities ($\ell = 10$). The larger the space capacity, the more challenging it is to accurately approximate the disaggregation probabilities (since there are more disaggregate states within the aggregate state).

Figure 4 considers for each scenario the average absolute error; it groups the scenarios according to common location of the bottleneck queue (i.e., queue with the highest traffic intensity). For a given scenario set (i.e., a given x-value in Figure 4), the only difference in the three scenarios is the location of the bottleneck queue; all other scenario parameters are common. The circles denote the scenarios where all three queues have common traffic intensity, and hence they are all considered bottlenecks. The crosses (respectively, squares) denote the scenarios where the bottleneck queue is the most upstream (respectively, most downstream) queue.

Figure 4 shows that for eight out of the nine sets of scenarios (i.e., all but set 6), the smallest errors are obtained when the bottleneck queue is the most upstream queue only. This can be explained as follows. When the bottleneck is located upstream of the network, blocking (e.g., spillback) effects are not likely to occur further downstream, and hence the between-queue dependencies are not as intricate as if the bottleneck were located further downstream. Since the bottleneck effects are very difficult to describe and approximate analytically, upstream bottlenecks are the scenarios with the highest accuracy in the predictions.

For seven out of the nine sets of scenarios (i.e., all but sets 3 and 6), the largest errors are obtained when the bottleneck is located at the most downstream queue only. As described above, this leads to significant spillback effects, and hence intricate between-queue dependencies that are difficult to approximate analytically.

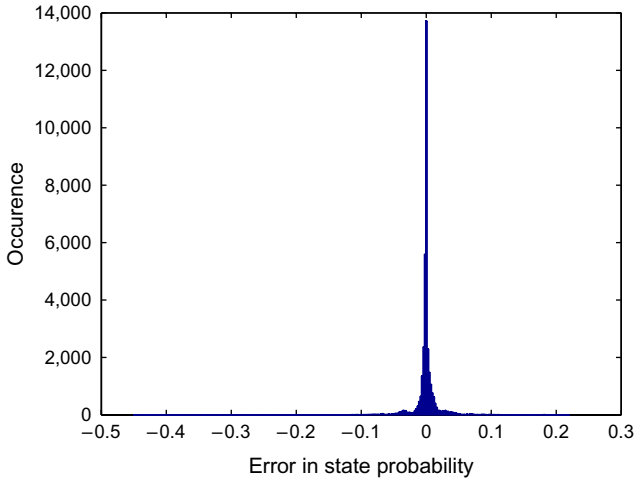
Figure 5 displays the errors for all scenarios, all state probabilities at all times. This considers a total of 36,450 probabilities, with an average absolute error of 0.0095. Across all scenarios the average runtime for the analytical method is 13.7 seconds, with a standard deviation of 0.74 seconds. All scenarios were run on a 1.7 GHz Intel Core i5 processor and 4 GB RAM.

We now consider an eight-queue tandem network with the scenario defined in Table 3. All queues have a common service rate $\mu = 10$. This leads to a network with increasing congestion as the queue index

Table 3. Eight-Queue Network Scenario

Queue i	1	2	3	4	5	6	7	8
γ_i	4	0	1	1	0	2	0	1
ℓ_i	25	10	25	10	25	10	25	10

Figure 5. (Color online) Histogram of the Errors for All 27 Scenarios, for All State Probabilities at All Times



increases. The traffic intensities of the queues increase from 0.4 to 0.9.

The proposed analytical approach decomposes an eight-queue network into six overlapping subnetworks. Each plot of Figure 6 considers the probabilities obtained by both the analytical model (blue circles), and the simulation model (red crosses with their corresponding 95% confidence intervals). The probabilities of all joint states of all subnetworks are displayed. Each plot considers a different time, going from time $t = 10$ in the upper plot, to $t = 20$, 30, 40, and 50 in the lower plots. Figure 6 shows that across time and across all subnetworks the analytical approach yields accurate approximations.

Figure 7. (Color online) Histogram of the Errors for All State Probabilities at All Times for the 8-Queue Network

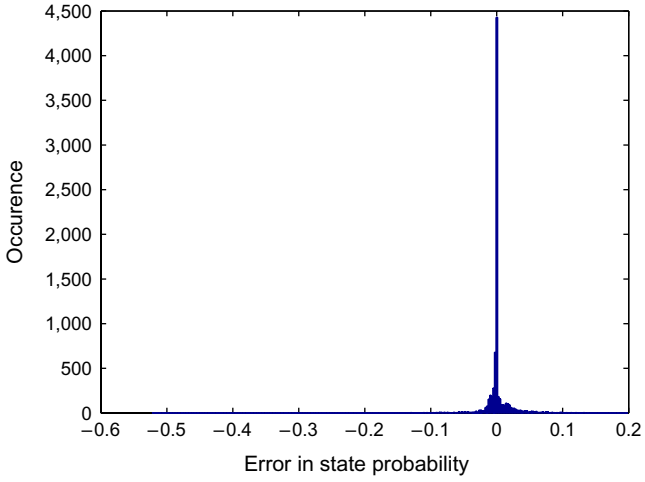


Figure 7 displays the errors for all state probabilities at all times. This considers a total of 8,100 probabilities, with an average absolute error of 0.0105. The runtime for the analytical method is 10.1 minutes.

We now consider a tandem network with 25 queues. The queues with even indices have $\ell_i = 10$, those with odd indices have $\ell_i = 25$. For all queues $\mu = 10$. The only nonzero external arrival rates are $\gamma_1 = 2$, $\gamma_{11} = 2$, $\gamma_{17} = 3$, and $\gamma_{21} = 2$. This leads to a network with increasing traffic intensity as the queue index increases, the traffic intensities vary from 0.2 to 0.9.

The analytical method decomposes the 25-queue network into 23 subnetworks. Figure 8 displays five

Figure 6. (Color online) State Probabilities for All States of All Subnetworks in the 8-Queue Network

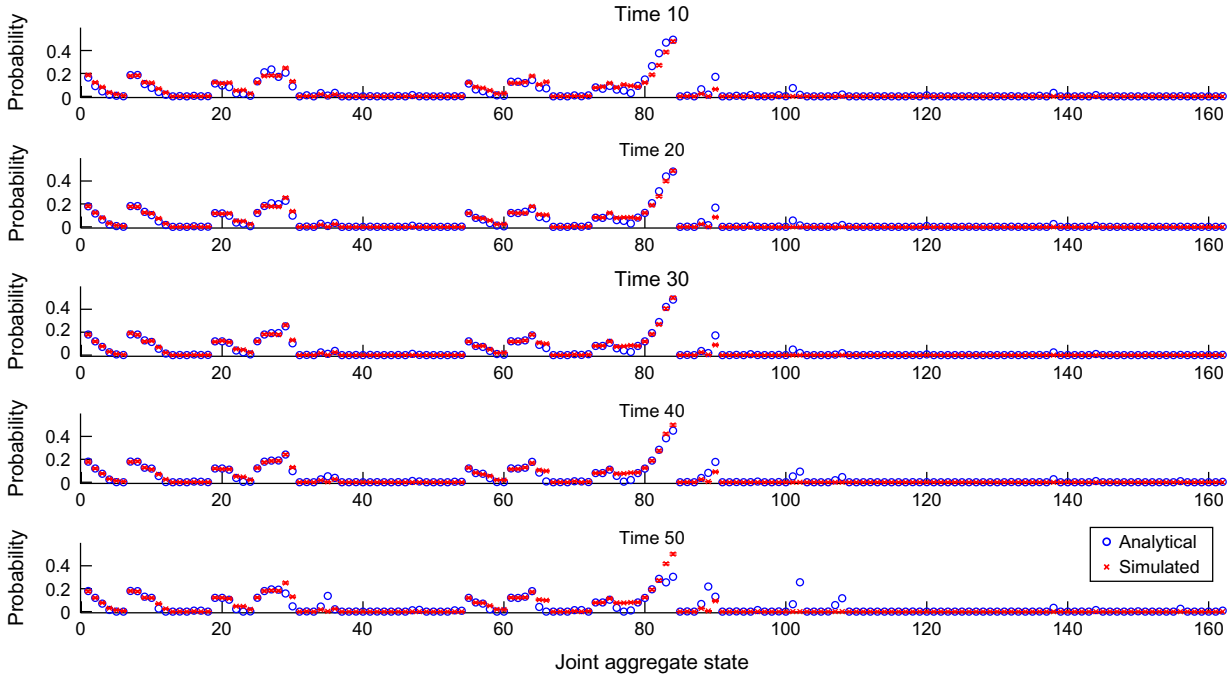
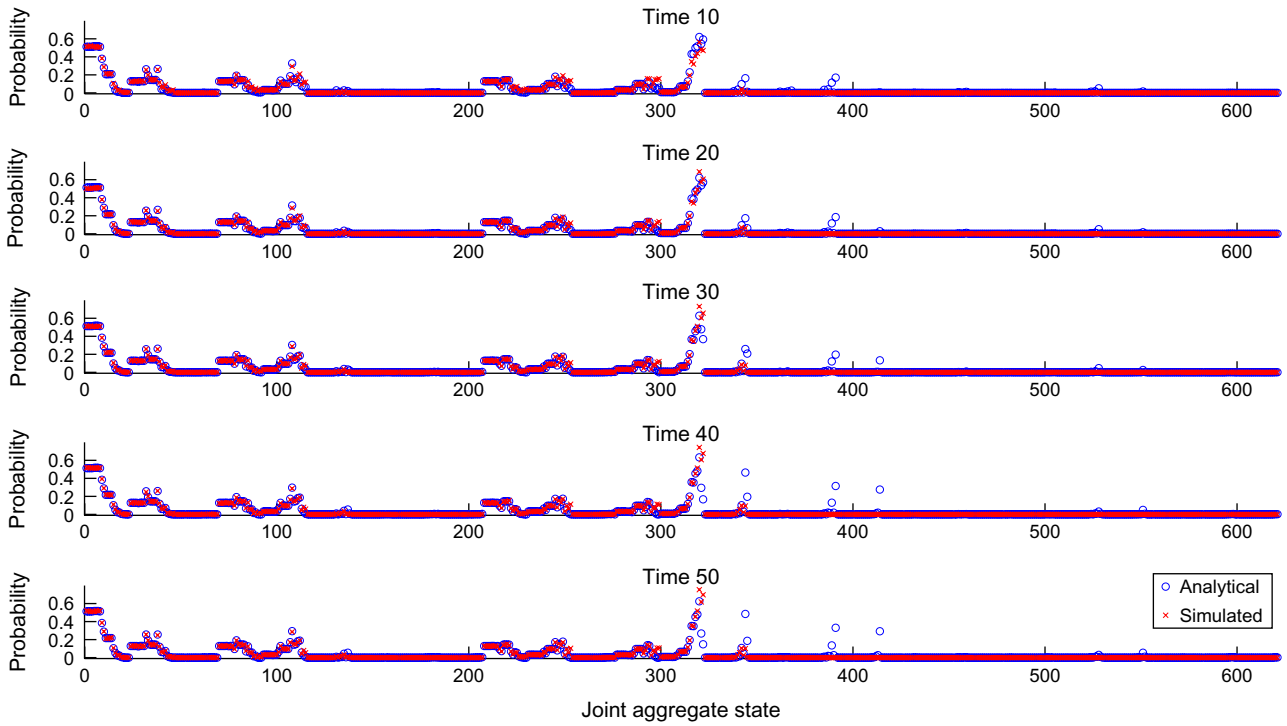
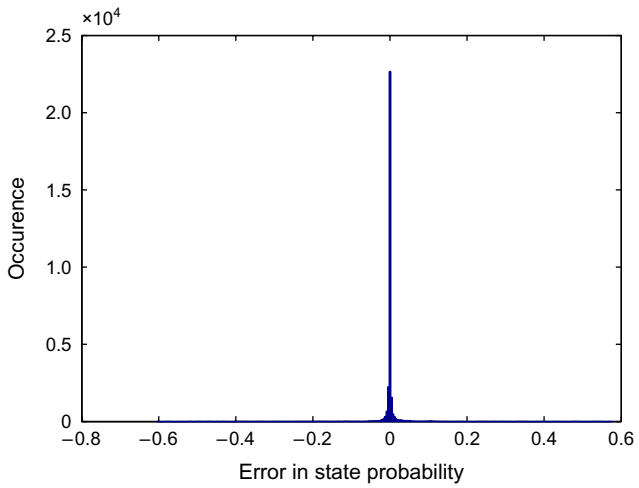


Figure 8. (Color online) State Probabilities for All States of All Subnetworks in the 25-Queue Network



plots, each plot considers a given time: $t = \{10, 20, 30, 40, 50\}$. Each plot displays the analytical (blue circles) and the simulated estimate (red crosses with their corresponding 95% confidence intervals) of the aggregate state probability, for all feasible aggregate states. Overall the proposed method provides a good approximation to the aggregate state probabilities. The corresponding histogram that considers the errors of all states at all times is displayed in Figure 9. Figure 9 considers a total of 31,050 state probabilities. The average absolute error is 0.0079. The runtime for the analytical method is 23.4 minutes.

Figure 9. (Color online) Histogram of the Errors for All State Probabilities at All Times for the 25-Queue Network



4. Urban Traffic Signal Control

This section considers an urban traffic signal control problem, and studies the added value of accounting for both transient and joint distributional information. We compare the performance of the signal plans proposed by (i) our proposed transient joint method, (ii) the stationary joint method of Osorio and Wang (2017), and (iii) a stationary marginal model, which approximates the (disaggregate) marginal queue-length distributions. The latter model is formulated in Osorio and Bierlaire (2009) and Osorio (2010, Chapter 4), its formulation for an urban network is given in Online Appendix B. Methods (i) and (ii) both consider subnetworks with three queues; method (i) considers a time-dependent description of between-queue dependencies, whereas method (ii) considers a stationary analysis. Hence, their comparison gives insights on the added value of accounting for the dynamics of between-queue dependencies. The comparison of methods (i) and (iii) gives insights on the added value of providing both a dynamic and a higher-order description of between-queue dependency. The performance of the signal plans proposed by the different models are evaluated by a microscopic stochastic urban traffic simulation model implemented in Aimsun version 6.1 (TSS 2011). Additional details regarding the simulation model can be found in Yamani (2013).

4.1. Road Network

The road network (see Figure 10) consists of 20 single-lane roads and four intersections, each with two en-

Figure 10. (Color online) Road Network

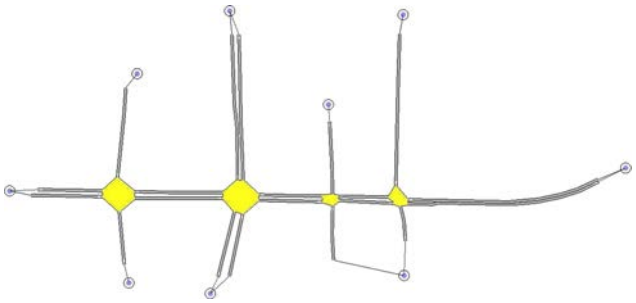
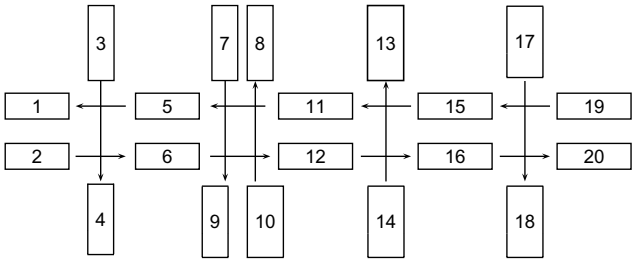


Figure 11. Queueing Network



dogenous 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 Figure 10). The queueing representation along with corresponding link/queue indices is displayed in Figure 11. In Figure 11 the queues are represented with rectangles.

We consider a medium demand and a high demand scenario (see Table 4). In Table 4, the indices in the first row correspond to link/queue indices as defined in Figure 11. We assume an initially empty network, and consider a time interval of 75 minutes.

4.2. Queueing Network

Let us describe how the road network is modeled as a queueing network. The below approach has been successfully used in past work that uses queueing-theoretic models of road transportation (Osorio and Bierlaire 2013).

All roads of the considered network are single-lane roads, each lane is modeled as one queue. The space capacity of a queue is given by

$$\ell_i = \lfloor (l_i + d_2)/(d_1 + d_2) \rfloor, \tag{27}$$

Table 4. Demand in Vehicles per Hour for the Two Demand Scenarios

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

Table 5. External Arrival Rates for Each Queue for the Two Demand Scenarios

	γ_{19}	γ_2	γ_3	γ_7	γ_{10}	γ_{14}	γ_{17}
Medium demand	700	700	100	600	600	100	100
High demand	900	900	100	600	600	200	200

where l_i is the length of lane i in meters, d_1 is the average vehicle length (set to four meters), and d_2 is the minimal intervehicle distance (set to one meter). The fraction is rounded down to the nearest integer. This expression for the space capacity follows similar ideas than those in Heidemann (1996) and Van Woensel and Vandaele (2007), where each road is divided into segments of length $1/k_{\text{jam}}$, where k_{jam} is the jam density of the lane. Hence, $1/k_{\text{jam}}$ represents the minimal distance that an average vehicle occupies.

The routing probability from queue i to queue j, denoted p_{ij}, is derived from turning probabilities. Based on Figure 11 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.

The external arrival rates of each queue, γ_i, are given by the origin-destination matrix of Table 4, and stated for each queue in Table 5. Queues not included in Table 5 have an external arrival rate of zero.

The service rate of a queue is defined as the downstream flow capacity of the underlying lane. For nonsignalized lanes, the service rate is equal to the saturation rate, s (set to 1,800 vehicles per hour). For signalized lanes, the service rate is given by

$$\mu_i = g_i s, \tag{28}$$

where g_i represents the total green split of queue i (i.e., ratio of total green time to intersection cycle time).

The subnetworks of the joint models (transient and stationary) are as follows. The cross streets (northbound and southbound) are modeled individually (i.e., they constitute singleton subnetworks). The links of the westbound and eastbound arterial are modeled jointly, i.e., the paths are decomposed into three-queue subnetworks. In other words, the subnetworks 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 parentheses are queue indices.

4.3. Problem Formulation

We consider a traffic signal control problem. For a review of traffic signal control terminology and formulations, we refer the reader to Appendix A of Osorio (2010) or to Lin (2011). The signal control problem that we consider is known as a fixed-time (also called time of day or pretimed) control strategy. For a given intersection and a given time interval (e.g., evening peak

period), a fixed-time signal plan is a cyclic (i.e., periodic) plan that is repeated throughout the time interval. The duration of the cycle is the time required to complete one sequence of signals. The sequence may contain all-red periods, where all streams have red indications, as well as stages with fixed durations (e.g., for safety reasons). The sum of the all-red periods and the fixed periods is called the fixed cycle time. Note that there has been interesting recent research for other families of traffic-responsive signal control problems (Varaiya 2013; Gregoire et al. 2015; Gayah, Gao, and Nagle 2014; He et al. 2014).

In this paper, the decision variables are the endogenous green splits (i.e., normalized green times) of each intersection. All other traditional control variables (e.g., cycle times, offsets, stage structure) are assumed fixed. The signal plans of all intersections are determined simultaneously.

To formulate this problem we introduce the following notation:

$[t_0, t_1]$ time interval of interest;
δ time step;
c_i cycle time of intersection i;
d_i fixed cycle time of intersection i;
e_l ratio of fixed green time to cycle time of signalized lane l;
s saturation flow rate [veh/h];
$x(j)$ green split of phase j;
x_L vector of minimal green splits;
y endogenous queueing model variables;
u exogenous queueing model parameters;
$\mathcal{I}$ set of intersection indices;
$\mathcal{L}$ set of indices of the signalized lanes;
$\mathcal{P}_I(i)$ set of phase indices of intersection i;
$\mathcal{P}_L(l)$ set of phase indices of lane l.

The problem is formulated as follows:

$$\min_x \quad g(x, y; u, t_0, t_1) \tag{29}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{P}_I(i)} x(j) = \frac{c_i - d_i}{c_i}, \quad \forall\, i \in \mathcal{I}, \tag{30}$$

$$\mu_l - \sum_{j\in\mathcal{P}_L(l)} x(j)s = e_l s, \quad \forall\, l \in \mathcal{L}, \tag{31}$$

$$h(y; u, t_0, t_1) = 0, \tag{32}$$

$$y \geq 0, \tag{33}$$

$$x \geq x_L. \tag{34}$$

The decision vector x is the vector of green splits for each phase. Constraints (30) relate, for each intersection i, its available cycle time to its endogenous phases. Constraints (31) relate the service rate (i.e., link flow capacity) of a signalized link to the saturation flow s (set to 1,800 vehicles per hour) and to its total green time. Equation (32) represents the traffic model, which 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). In the case of the proposed transient model, u includes the time step δ and initial probability distributions. The endogenous queueing variables are subject to positivity constraints (33). Green splits have lower bounds (Equation (34)), which are set to four seconds following the transportation norms (VSS 1992). The objective function $g(x, y; u, t_0, t_1)$ represents the expected trip travel time during $[t_0, t_1]$.

For the proposed transient model, the objective function is given by

$$g(x, y; u, t_0, t_1) = \frac{1}{K}\sum_{k=1}^{K} g^k(x, y; u, t_0, t_1), \tag{35}$$

where K is the total number of discrete time intervals, and g^k represents the expected travel time during time interval k. The latter is obtained by applying Little's law at the end of the time interval (Little 1961, 2011)

$$g^k(x, y; u, t_0, t_1) = \frac{\sum_{i=1}^{I} E[N_i^k]}{\sum_{i=1}^{I} \gamma_i p^k_{A_i\neq 2}(\delta)}, \tag{36}$$

where the summations consider all I queues in the network, and $E[N_i^k]$ represents the expected number of vehicles in queue i at the end of time interval k

$$E[N_i^k] = \sum_{n=0}^{\ell_i} n p^k_{N_i=n}(\delta). \tag{37}$$

The disaggregate distribution for queue i at time interval k is obtained by solving the below system of equations to obtain λ_i^k and μ_i^k, which then fully define the disaggregate distribution according to the System of Equations (15)

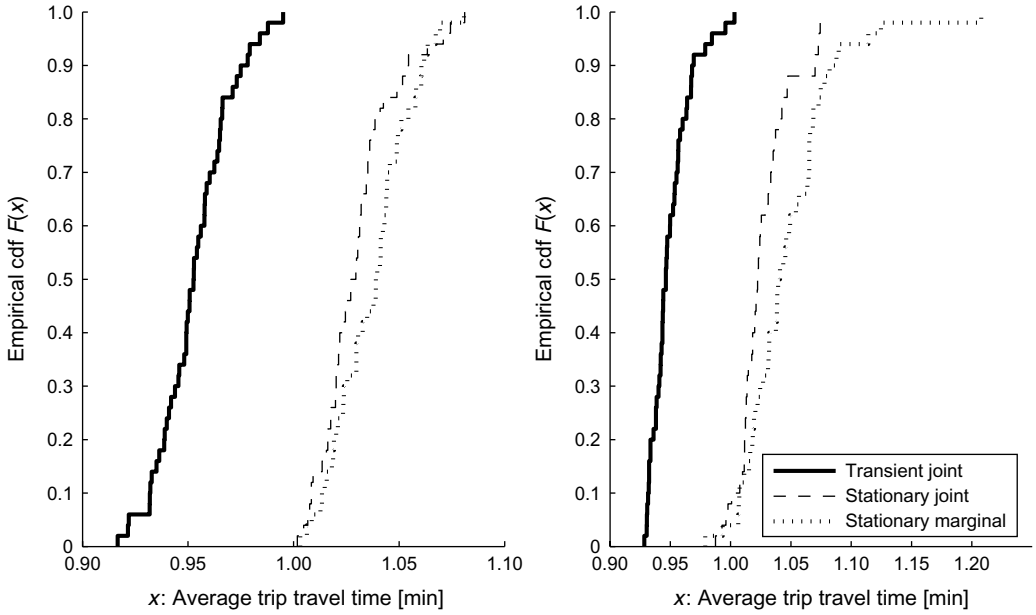
$$\begin{cases} p^k_{A_i=0}(\delta) = f_D(0, \delta, \lambda_i^k, \mu_i^k, p^{k-1}_{N_i}(\delta)), & \text{(38a)} \\ p^k_{A_i=2}(\delta) = f_D(\ell_i, \delta, \lambda_i^k, \mu_i^k, p^{k-1}_{N_i}(\delta)). & \text{(38b)} \end{cases}$$

For each queue i, its aggregate distribution $p^k_{A_i}$ is derived from the analysis of subnetwork i.

4.4. Implementation Notes

For the proposed model we set the time step $\delta = 0.1$. The signal control problem is solved using the *active-set* algorithm of the *fmincon* solver of MATLAB (Mathworks 2011) with constraint and objective function tolerance of 10^{-6} and 10^{-3}, respectively. The stationary joint model as well as our proposed transient joint model both use the plan considered optimal by the stationary marginal model (Osorio 2010, Chapter 4) as their initial signal plan. More details on how the algorithms are initialized are included in Osorio and Wang (2017, Section 4.3). The runtime to solve the optimization problem using the transient joint method is 28 hours.

Figure 12. The Left (Respectively, Right) Plot Displays the cdfs of the Average Trip Travel Time Considering the Medium (Respectively, High) Demand Scenario



4.5. Results

The performance of a given signal plan is evaluated by embedding the signal plan within a microscopic stochastic traffic simulator of the network depicted in Figure 10 and running 50 simulation replications. For each replication, we obtain a realization of the objective function: the average trip travel time (ATTT). For each signal plan, we use the 50 simulated observations of the ATTT to construct a cumulative distribution function (cdf). Figure 12 displays several cdf curves. The x-axis displays the ATTT. For a given x, the y-axis displays the proportion of simulation replications (out of the 50 replications) that have ATTT values smaller than x. Hence, the more the cdf curves are shifted to the left, the higher the proportion of small ATTT values.

The left (respectively, right) plot of Figure 12 displays the results considering the medium (respectively, high) demand scenario. Each plot contains three cdf curves. The solid curve corresponds to the signal plan derived by our proposed transient joint model. The dashed (respectively, dotted) curve corresponds to the plan of the stationary joint (respectively, stationary marginal) model. For both demand scenarios, the proposed approach significantly outperforms the other two approaches. It outperforms the stationary joint approach, which shows the added value of accounting for transient information. Both joint approaches (transient joint and stationary joint) outperform the marginal approach, showing the added value of providing a higher-order (i.e., beyond first-order) description of the between-queue dependency.

We test the hypothesis that the expected trip travel time derived from the joint transient model is equal to that derived by the joint stationary model by conducting a paired t-test. Denoting the sample mean of the paired differences as $\hat{Y}$, the standard deviation as $\hat{s}$, and the number of observations as O, a paired t-statistic is given by Hogg and Tanis (2006, p. 486): $\sqrt{O}\hat{Y}/\hat{s}$. For both the medium and the high demand scenario, the mean of the paired differences (i.e., difference between the average trip travel time given by the joint stationary model and that given by the joint transient model) is approximately 0.077 minutes. The standard deviation of the paired differences is approximately 0.024 (respectively, 0.029) minutes for the medium (respectively, high) demand scenario. Thus, for the 50 observations, the test statistic is 22.32 (respectively, 19.06) for the medium (respectively, high) demand scenario. The null hypothesis is rejected for both demand scenarios, as the critical value, $t_{0.01}(49) = 2.405$, is less than the value of the test statistic. The improvement in average trip travel time is statistically significant.

5. Conclusions

This paper proposes an analytical, tractable, and scalable technique that approximates the transient aggregate joint queue-length distribution of a finite (space) capacity tandem Markovian network. The complexity of the proposed method is linear, rather than exponential, in the number of queues and is independent of the queue space capacities, making it a suitable approach for the analysis of large-scale networks.

The analytical approximations of the aggregate joint distributions are validated versus estimates obtained via discrete-event simulation of a queueing network. The validation scenarios consider various congested

networks. The analytical approximations are very accurate. The model is then used to address an urban traffic signal control problem. The proposed model yields signal plans that significantly outperform those derived by a stationary joint model, as well as those derived by a stationary marginal model. This shows the added value of using a higher-order description of the spatial-temporal between-link dependencies to devise traffic management strategies for congested urban networks.

Extensions of this work include its formulation for a general topology network. Additionally, it can be used to improve the computational efficiency of dynamic simulation-based optimization algorithms following the frameworks in Osorio and Bierlaire (2013) and Osorio and Chong (2015).

References

Balsamo S, de Nitto Personé V, Onvural R (2001) *Analysis of Queueing Networks with Blocking*, Internat. Series Oper. Res. Management Sci., Vol. 31 (Kluwer Academic Publishers, Boston).

Barceló J (2010) *Fundamentals of Traffic Simulation*, Internat. Series Oper. Res. Management Sci., Vol. 145 (Springer-Verlag, New York).

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

Chong L, Osorio C (2017) A simulation-based optimization algorithm for dynamic large-scale urban transportation problems. *Transportation Sci.*, ePub ahead of print July 19, http://dx.doi.org/10.1287/trsc.2016.0717.

Cohen JW (1982) Some variants of the single server queue. Cohen JW, ed. *The Single Server Queue*, Applied Math. Mechanics, Vol. 8 (North-Holland, Amsterdam).

Coleman TF, Li Y (1994) On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds. *Math. Programming* 67(2):189–224.

Coleman TF, Li Y (1996) An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.* 6(2):418–445.

Department of Transportation (2008) Transportation vision for 2030. Technical report, U.S. Department of Transportation (DOT), Research and Innovative Technology Administration, Washington, DC.

Durrett R (1999) *Essentials of Stochastic Processes*, Springer Texts Statistics (Springer-Verlag, New York).

Flötteröd G, Osorio C (2013) Approximation of time-dependent multi-dimensional queue-length distributions. *Proc. Triennial Symp. Transportation Anal. (TRISTAN).*

Gayah VV, Gao XS, Nagle AS (2014) On the impacts of locally adaptive signal control on urban network stability and the macroscopic fundamental diagram. *Transportation Res. Part B* 70: 255–268.

Gregoire J, Qian X, Frazzoli E, de La Fortelle A, Wongpiromsarn T (2015) Capacity-aware back-pressure traffic signal control. *IEEE Trans. Control Networked Systems* 2(2):164–173.

Griffiths JD, Leonenko GM, Williams JE (2008) Approximation to the transient solution of the M/Ek/1 queue. *INFORMS J. Comput.* 20(4):510–515.

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

He Q, Head KL, Ding J (2014) Multi-modal traffic signal control with priority, signal actuation and coordination. *Transportation Res. Part C* 46:65–82.

Heidemann D (1996) A queueing theory approach to speed-flow-density relationships. *Proc. 13th Internat. Sympos. Transportation Traffic Theory*, 103–118.

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

Hogg R, Tanis E (2006) *Probability and Statistical Inference*, 7th ed. (Pearson Education, Upper Saddle River, NJ).

Kaczynski WH, Leemis LM, Drew JH (2012) Transient queueing analysis. *INFORMS J. Comput.* 24(1):10–28.

Larson R, Odoni A (1981) *Urban Operations Research* (Prentice-Hall, Englewood Cliffs, NJ).

Li J (2005) Overlapping decomposition: A system-theoretic method for modeling and analysis of complex manufacturing systems. *IEEE Trans. Automation Sci. Engrg.* 2(1):40–53.

Lighthill M, Witham J (1955) On kinematic waves II. A theory of traffic flow on long crowded roads. *Proc. Roy. Soc. London Series A* 229:317–345.

Lin S (2011) Efficient model predictive control for large-scale urban traffic networks. Unpublished doctoral thesis, Delft University of Technology, Delft, Netherlands.

Little JDC (1961) A proof for the queuing formula: $L = \lambda W$. *Oper. Res.* 9(3):383–387.

Little JDC (2011) Little's law as viewed on its 50th anniversary. *Oper. Res.* 59(3):536–549.

Mathworks (2011) Global optimization toolbox: User's guide. User's Guide Matlab version 2011b.

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

Meier P (2007) Simulation d'un réseau de files d'attente à capacités finies. Technical report, ROSO Chair of Operations Research SO, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Morse P (1958) *Queues, Inventories and Maintenance*: The Analysis of Operational Systems with Variable Demand and Supply (Wiley, New York).

Odoni AR, Roth E (1983) An empirical investigation of the transient behavior of stationary queueing systems. *Oper. Res.* 31(3): 432–455.

Osorio C (2010) Mitigating network congestion: Analytical models, optimization methods and their applications. Unpublished doctoral thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

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 (2013) A simulation-based optimization framework for urban transportation problems. *Oper. Res.* 61(6): 1333–1345.

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

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

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

Osorio C, Selvam K (2017) Simulation-based optimization: Achieving computational efficiency through the use of multiple simulators. *Transportation Sci.* 51(2):395–411.

Osorio C, Wang C (2017) On the analytical approximation of joint aggregate queue-length distributions for traffic networks: A stationary finite capacity Markovian network approach. *Transportation Res. Part B* 95:305–339.

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

Osorio C, Chen X, Marsico M, Talas M, Gao J, Zhang S (2015) Reducing gridlock probabilities via simulation-based signal control. *Transportation Res. Procedia, 4th Internat. Sympos. Transport Simulation (ISTS)*, Vol. 6, 101–110.

Peterson MD, Bertsimas DJ, Odoni AR (1995a) Decomposition algorithms for analyzing transient phenomena in multiclass queueing networks in air transportation. *Oper. Res.* 43(6):995–1011.

Peterson MD, Bertsimas DJ, Odoni AR (1995b) Models and algorithms for transient queueing congestion at airports. *Management Sci.* 41(8):1279–1295.

Reibman A (1991) A splitting technique for Markov chain transient solution. Stewart WJ, ed. *Numerical Solution of Markov Chains* (Marcel Dekker, New York), 373–400.

Rice JA (1994) *Mathematical Statistics and Data Analysis* (Duxbury Press, Belmont, CA).

Richards PI (1956) Shock waves on highways. *Oper. Res.* 4(1):42–51.

Schmidt LC, 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. Stewart W, ed. *Numerical Solutions of Markov Chains* (Marcel Dekker, New York), 63–88.

Schweitzer PJ (1984) Aggregation methods for large Markov chains. Iazeolla G, Courtois PJ, Hordijk A, eds. *Mathematical Computer Performance and Reliability* (North-Holland, Amsterdam), 275–286.

Sharma OP, Gupta UC (1982) Transient behavior of an M/M/1/N queue. *Stochastic Processes Their Appl.* 13(3):327–331.

Sharma OP, Shobha B (1988) Transient behaviour of a double-channel Markovian queue with limited waiting space. *Queueing Systems* 3(1):89–96.

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

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

Texas Transportation Institute (2012) 2012 Urban mobility report. Technical report, Texas Transportation Institute (TTI), Texas A&M University System, College Station.

Transport for London (2010) Traffic modelling guidelines. Version 3.0. Technical report, Transport for London (TfL), London.

TSS (2011) AIMSUN 6.1 Microsimulator Users 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.

Varaiya P (2013) Max pressure control of a network of signalized intersections. *Transportation Res. Part C* 36:177–195.

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.

Whitt W (1999) Decomposition approximations for time-dependent Markovian queueing networks. *Oper. Res. Lett.* 24(3):97–103.

Yamani JH (2013) Approximation of the transient joint queue-length distribution in tandem networks. Unpublished Master's thesis, Massachusetts Institute of Technology, Cambridge.

Zhang L, Wang C, Arinez J, Biller S (2013) Transient analysis of Bernoulli serial lines: Performance evaluation and system-theoretic properties. *IIE Trans.* 45(5):528–543.