Analytical approximation of transient joint queue-length distributions of a finite capacity queueing network

Gunnar Flötteröd * Carolina Osorio †

March 29, 2013

Problem statements

This work contributes to the modeling of analytical tractable joint distributions of time-dependent finite capacity Markovian queueing network states. The model is motivated by an interest to analyse network effects in congested urban road networks. The development of analytical, probabilistic and time-dependent models for urban traffic is seldom. In urban transportation, as in most fields, time-dependent (i.e., transient) analysis is most often carried out by either the use of (i) deterministic analytical models (such as the fundamental kinematic wave model (Lighthill and Witham; 1955; Richards; 1956)) or (ii) simulation-based models (for a review, see Barceló (2010)), which when stochastic can provide distributional estimates that go beyond first-order information (i.e., expectations). There is currently a lack of analytical transient techniques for finite capacity Markovian queueing networks that account for spatial-temporal dependencies, and even more a lack of tractable (i.e., computationally efficient) techniques. This paper proposes a tractable analytical approximation of transient multi-dimensional queue-length distributions.

In the broader field of transportation (all modes considered), few analytical probabilistic and time-dependent techniques have been developed (for a single queue see Heidemann (2001); Peterson et al. (1995b); for networks of queues see Osorio and Flötteröd (2012); Osorio et al. (2011); Gupta (2011); Peterson et al. (1995a); Odoni and Roth (1983)). This work contributes to the probabilistic analytical modeling of urban traffic. The proposed network model can be coupled with detailed link models of urban traffic (Osorio and Flötteröd; 2012; Osorio et al.; 2011), which are also derived based on queueing network theory and are consistent with the mainstream deterministic traffic models. The combination of such models can describe in a tractable and detailed manner both the within-link and the between-link dynamics, leading to a detailed probabilistic description of the spatial-temporal evolution of urban congestion.

Methodology

Consider a network of queues in an arbitrary topology, consisting of single server queues. External arrivals to a queue arise following an independent Poisson process, and service times are independently, identically distributed exponential random variables. Routing is probabilistic.

The following events are possible in the network: (i) arrivals from outside the network into a queue, (ii) departures from a queue out of the network, (iii) transitions from an upstream queue into a downstream

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

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

queue. The rates at which these events occur may be time-dependent. All queues have finite capacities, enabling the modeling of spill-back effects.

Denoting by $\mathbf{N} = \mathbf{N}(t)$ the random vector of all queue states in the network at real-valued time t, the dynamics of the joint distribution of $\mathbf{N}$ are guided by the following linear system of differential equations (Reibman; 1991):

$$\frac{d}{dt}P(\mathbf{N} = \mathbf{y}) \quad = \quad \sum_{\mathbf{x}} t_{\mathbf{x}}^{\mathbf{y}} P(\mathbf{N} = \mathbf{x}) \tag{1}$$

where $\frac{d}{dt}P$ is the time derivative of P, $\mathbf{x} = (x_i)$ and $\mathbf{y} = (y_i)$ are both realizations of $\mathbf{N}$, and $t_{\mathbf{x}}^{\mathbf{y}}$ is the transition rate from state $\mathbf{x}$ into state $\mathbf{y}$.

This model becomes computationally intractable for non-trivial networks due to the exponential increase in the dimension of the state space as a function of the number of queues. This work hence proposes a transient decomposition technique that characterizes the high-dimensional joint distribution $P(\mathbf{N})$ in terms of lower-dimensional marginals over non-disjoint ("overlapping") subnetworks, where a given queue belongs to the subnetworks of both its upstream and its downstream queues. This decomposition is not expensive to compute and, as proven in the full paper, unique.

Let S be a subnetwork and ∂S be the set of all of its adjacent subnetworks. The dynamics of the state $\mathbf{N}_{S,\partial S}$ of all queues in S and ∂S only can then be expressed as follows:

$$\frac{d}{dt}P(\mathbf{N}_{S,\partial S} = (\mathbf{m}, \mathbf{s})) \quad = \quad \sum_{\mathbf{n}} \sum_{\mathbf{q}} t_{(\mathbf{n},\mathbf{q})}^{(\mathbf{m},\mathbf{s})}(S, \partial S) P(\mathbf{N}_{S,\partial S} = (\mathbf{n}, \mathbf{q})) \tag{2}$$

where $(\mathbf{n}, \mathbf{q})$ and $(\mathbf{m}, \mathbf{s})$ are elements of the combined state space of S and ∂S, with the respective first tuple element belonging to S and the second element belonging to ∂S, and $t_{(\mathbf{n},\mathbf{q})}^{(\mathbf{m},\mathbf{s})}(S, \partial S)$ is the transition rate from state $(\mathbf{n}, \mathbf{q})$ to $(\mathbf{m}, \mathbf{s})$ when only accounting for events possible in subnetwork S and ∂S.

Summing out the queue states $\mathbf{s}$ and rearranging terms results in

$$\begin{aligned} \frac{d}{dt}P(\mathbf{N}_S = \mathbf{m}) \quad &= \quad \sum_{\mathbf{s}} \sum_{\mathbf{n}} \sum_{\mathbf{q}} t_{(\mathbf{n},\mathbf{q})}^{(\mathbf{m},\mathbf{s})}(S, \partial S) P(\mathbf{N}_{S,\partial S} = (\mathbf{n}, \mathbf{q})) \\ &= \quad \sum_{\mathbf{n}} \left[\sum_{\mathbf{q}} \sum_{\mathbf{s}} t_{(\mathbf{n},\mathbf{q})}^{(\mathbf{m},\mathbf{s})}(S, \partial S) P(\mathbf{N}_{\partial S} = \mathbf{q} \mid \mathbf{N}_S = \mathbf{n}) \right] P(\mathbf{N}_S = \mathbf{n}) \end{aligned} \tag{3}$$

where $\mathbf{N}_S$ and $\mathbf{N}_{\partial S}$ are the random vectors of queue states only in subnetworks S and ∂S, respectively. It is noteworthy that the term in brackets functions like a state-dependent transition rate from subnetwork state $\mathbf{n}$ to subnetwork state $\mathbf{m}$. This – so far exact – expression is the basis for the proposed model.

The full article develops in detail a tractable approximation of these state-dependent transition rates. This approximation only requires computations based on subnetwork states, resulting in a model specification where the dynamics of all overlapping subnetworks are jointly determined, with each subnetwork only evaluating the instantaneous states of its neighbouring subnetworks.

Further, the article proves that if any two subnetwork distributions have identical marginals for a common set of queues at some point in time, this marginal will also be identical at all other points in time. This constitutes a key feature of the proposed decomposition approach: It maintains mutually consistent overlapping subnetwork distributions without any need to introduce supplementary distributional adjustments or constraints.

Illustrative results

Consider the network shown in Figure 1, consisting of eight single-lane roads. It is constructed in the same spirit as the network presented by Corthout et al. (2012). Two inflows compete for the network capacity,

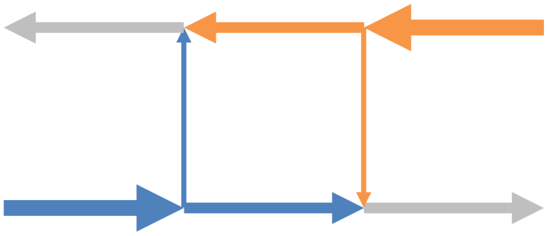

Figure 1: Conflicting network inflows

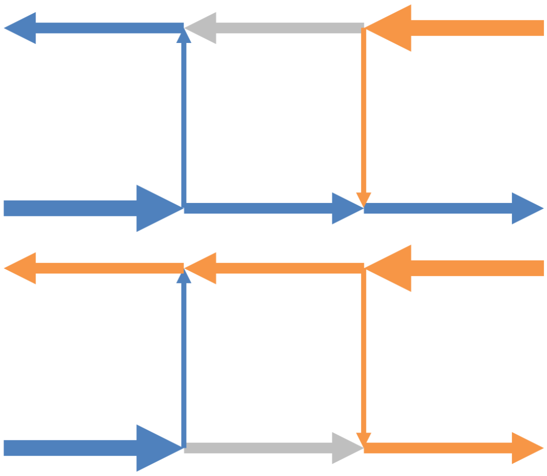

Figure 2: Multiple solutions

one entering at the south-western link (blue) and one entering at the north-eastern link (orange). The flows split at the intersections proportionally to the widths of the respective arrows. At the merge points, straight flows have priority over flows entering from the side.

A deterministic version of this model has two solutions, as shown in Figure 2. In the first (top) solution, the blue flow dominates the south-eastern merge, such that the orange flow spills back and cannot enter the north-western merge, where the blue flow can hence enter unhindered. Symmetrically, in the second (bottom) solution the orange flow dominates the north-western merge. The blue flow hence spills back, without entering the south-eastern merge, where the orange flow can hence pass unhindered.

In the presence of stochasticity, this system may oscillate between these two solutions, with the period of these oscillations depending on the network parameters. Indeed, it is possible to construct arbitrarily long periods of these oscillations, making a simulation-based analysis of this system extremely cumbersome. To give an example, Figure 3 shows the occupancy of the middle link in the lower road stretch, going from west to east, over simulation time. It fluctuates between being almost completely full (containing 9 or 10 vehicles, meaning that the blue flow spills back) and a more distributed state corresponding to maximum flow throughput (meaning that the blue flow has taken over).

Figure 4 shows an estimation of the time-dependent expected occupancy on the same link as computed by the proposed analytical model. The model captures two important effects. First, its transient nature allows to identify an initial overshoot in expected occupancy, resulting from the fact one starts with an empty network, such that initially the demand can enter and occupy the network unhindered from both sides. Only after a while, spillback effects occur. Second, it takes more than 300 time steps (5 minutes) to attain stationarity in this system, due to its oscillatory nature. The full article investigates in greater detail how the proposed model approximates multi-variate network distributions of this type.

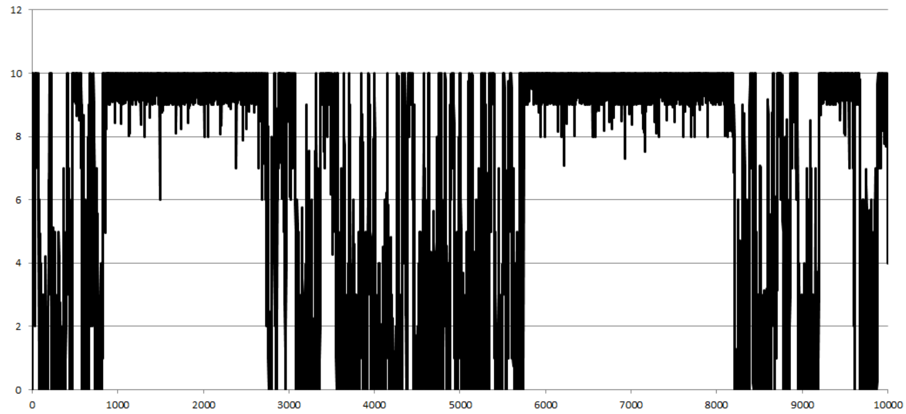

Figure 3: Fluctuations of link occupancy over time

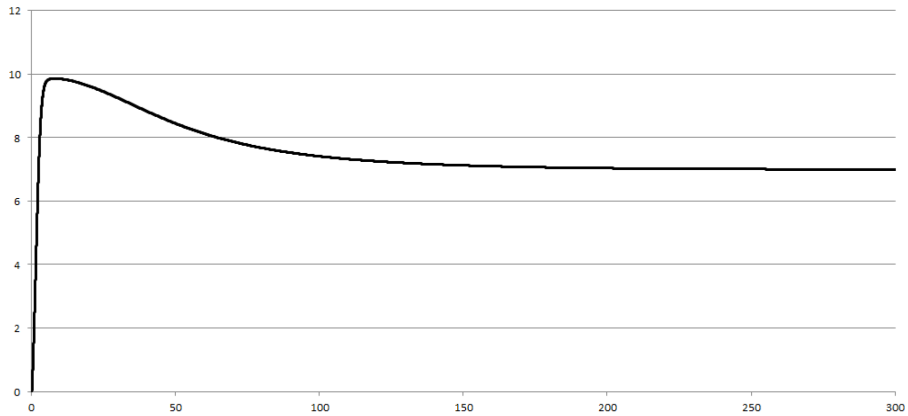

Figure 4: Analytically computed expectation of link occupancy over time

References

Barceló, J. (2010). *Fundamentals of traffic simulation*, Vol. 145 of *International Series in Operations Research and Management Science*, Springer, New York, USA.

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

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

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

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

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

Osorio, C. and Flötteröd, G. (2012). Capturing dependency among link boundaries in a stochastic network loading model, *International Symposium on Dynamic Traffic Assignment (DTA)*, Martha's Vineyard, USA.

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

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

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

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

Richards, P. I. (1956). Shock waves on highways, *Operations Research* **4**(1): 42–51.