A scalable algorithm for the control of congested urban networks with intricate traffic patterns: New York City case studies

Carolina Osorio[*] Xiao Chen[†] Jingqin Gao[‡] Mohamad Talas[§] Michael Marsico[¶]

1 Introduction

This paper focuses on the control of large-scale congested urban networks with intricate traffic dynamics. We consider networks with the following properties. They are highly and uniformly congested. Congestion arises in all directions of travel. Hence, congestion patterns do not reveal a clear hierarchy or priority between intersections or between links. The links are configured in a grid topology. This allows for intricate traveler behavior, such as high-dimensional route choice alternatives. This makes the prediction of how traffic (i.e., travelers) will respond to changes in the network supply a greater challenge. Most links of the network are short links, which are prone to the occurrence of spillbacks, also known as spillovers, and contribute to a rapid spatial propagation of congestion. The networks have multi-modal traffic (e.g., cars, transit). Additionally, the case studies of this paper consider networks with heavy pedestrian traffic. Hence, the design of control strategies for vehicular traffic (e.g., traffic signals) is constrained by non-negligible pedestrian crossing times, and this at every intersection.

For such networks, the main shortcoming of existing signal control methods is the lack of a detailed modeling, or estimation, of between-link interactions, which enables an accurate description of the spatial propagation of congestion. Most signal control strategies do not account for between-link (i.e., between-queue) interactions. They are based on the use of vertical queues, also called

[*] Massachusetts Institute of Technology, Cambridge, MA, USA
[†] School of Highway, Chang'an University, Xi'an, China
[‡] New York City Department of Transportation, New York, NY, USA
[§] New York City Department of Transportation, New York, NY, USA
[¶] New York City Department of Transportation, New York, NY, USA

point queues, that do not account for the spatial propagation of vehicular queues. Thus, they do not describe phenomena such as spillbacks. They are appropriate for low to moderate levels of congestion, yet are unsuitable for highly congested networks (Papageorgiou et al.; 2003; Abu-Lebdeh and Benekohal; 2003).

For congested networks, mitigating the spatial propagation of congestion, as well as the occurrence of spillbacks, is recognized as a major goal (Chow and Lo; 2007; Abu-Lebdeh and Benekohal; 2003). For instance, in the New York City case study of Section 3, the occurrence of spillbacks affects the access to, and the egress from, the Queensboro Bridge, leading to significant impacts on the traffic throughput between Manhattan and Queens. The work of Geroliminis and Skabardoni (2011); Skabardonis and Geroliminis (2008) emphasize the importance of controlling queue spillbacks: "when spillovers occur, the travel delay can increase by 50%-100% for short distance links between successive intersections." The design of algorithms to mitigate the spatial propagation of congestion requires models that can describe how links interact under congested conditions, such as the occurrence and impact of spillbacks.

Several signal control methods based on queue-length information have been proposed. Examples include Abu-Lebdeh and Benekohal (2003); Diakaki et al. (2002); Michalopoulos and Stephanopoulos (1977a,b). They are based on the use of macroscopic traffic models. The vast majority of the research in the field of macroscopic traffic modeling has focused on the development of link models, which describe the within-link propagation of congestion. Formulations of the between-link interactions in an urban environment (i.e., node models) are limited (Lebacque and Khoshyaran; 2005; Tampère et al.; 2011; Flötteröd and Rohde; 2011; Corthout et al.; 2012). This is arguably due to the difficulty of providing an analytical, let alone differentiable, description of node priorities. Existing macroscopic approaches therefore embed a highly simplified, and most often non-differentiable, description of between-link interactions. The lack of differentiability limits their use within traditional gradient-based algorithms for large-scale network optimization.

More recently, various signal control methods that rely on real-time queue-length estimates have been proposed (Gregoire et al.; 2014; Lioris et al.; 2014; Varaiya; 2013; Wongpiromsarn et al.; 2012). These are scalable algorithms. Unfortunately, most urban networks currently do not have sensors deployed to provide accurate queue-length measurements (Papageorgiou and Varaiya; 2009). The extension of these algorithms to model between-queue interactions in the absence of accurate queue-length estimates is a topic of ongoing research.

The challenges of developing macroscopic analytical, let alone differentiable, network models that provide a suitable description of between-link interactions while remaining sufficiently tractable for large-scale control have led to an increased focus on the use of simulation-based models. For a review, see Barceló (2010). In particular, high-resolution simulators (e.g., mesoscopic, microscopic) can provide a detailed description of the intricate between-queue interactions. This is because these simulators represent individual vehicles and embed a detailed description of the network supply (e.g., prevailing traffic management strategies). Hence, they yield a high-resolution description of intricate local traffic phenomena, such as spillbacks. Nonetheless, the computational inefficiency of these high-resolution models has mostly confined their use to what-if analysis (i.e., scenario-based analysis), such as in Bullock et al. (2004); Ben-Akiva et al. (2003). Their use within simulation-based optimization (SO) algorithms is limited (Osorio and Nanduri; 2015a,b; Osorio and Chong; 2015; Osorio and Selvam; 2016; Osorio and Bierlaire; 2013; Li et al.; 2010; Stevanovic et al.; 2009, 2008; Branke et al.; 2007; Yun and Park; 2006; Hale; 2005; Joshi et al.; 1995). These SO algorithms enable signal plans to be designed based on the use of simulators with a high-resolution description of between-link interactions.

In summary, the signal control literature recognizes the importance, for congested networks, of using models that provide a detailed description of between-link interactions. There is currently a lack of macroscopic traffic-theoretic node models suitable for the optimization of large-scale urban networks. This is a topic of active ongoing research in the field. On the other hand, high-resolution simulation-based models (e.g., mesoscopic, microscopic) can provide a detailed description of the between-link interactions.

This paper focuses on the design of SO algorithms that enable the use of high-resolution simulators for the optimization of large-scale urban networks with intricate congested traffic patterns. There is a need for SO algorithms that can identify solutions to transportation problems in a computationally efficient manner, i.e., algorithms that can identify points with improved performance within few simulation runs. Such algorithms enable an efficient use of these inefficient simulators. They are of particular importance and relevance for transportation practitioners, who typically address the optimization problems under tight computational budgets (i.e., few simulation runs are carried out).

In our past work, efficiency has been achieved by providing the SO algorithm with analytical structural problem-specific information derived from macroscopic traffic models. A more detailed description of how this is done is given in Section 2. Basically, the SO algorithm solves a series of subproblems which are analytically constrained by a macroscopic traffic model. The macroscopic model provides the algorithm with an analytical, differentiable and global (i.e., in the entire feasible region) description of the mapping between the decision variables and the unknown simulation-based objective function. This differs from traditional SO algorithms that treat the simulator as a black-box (Stevanovic et al.; 2008; Park et al.; 2009). For various types of optimization problems, we have shown that it is this analytical problem-specific information that allows the algorithm to achieve efficiency, i.e., to identify, within few simulation runs, points with improved performance. (Osorio et al.; 2016; Osorio and Nanduri; 2015a; Osorio and Chong; 2015; Osorio and Nanduri; 2015b; Chen et al.; 2012; Osorio and Bierlaire; 2013). As is discussed in Section 2, under tight computational budgets, SO algorithms that embed analytical information from macroscopic models outperform those that do not.

As the scale of the networks and the levels of congestion increase, so does the computational runtime of these simulators. For large-scale congested networks, there is a greater need for efficient SO algorithms. This paper proposes an SO algorithm that allows large-scale networks and high-dimensional SO problems to be efficiently addressed. As is detailed in Section 2, the family of SO algorithms considered solves at every iteration, a subproblem that is constrained by an analytical macroscopic traffic model. The formulation of SO algorithms that remain efficient for larger-scale networks and higher-dimensional problems requires the formulation of analytical traffic models that provide a good approximation of the unknown simulation-based objective function, while also being differentiable, scalable and efficient to evaluate. This paper formulates such a model. The contributions of this paper are the following.

- **Analytical structural information.** In order to design scalable SO algorithms, Section 3 addresses the following question: what is the key analytical information, provided by the macroscopic models to the SO algorithms, that contributes to their computational efficiency? The main insights obtained are the following. Although the simulator embeds a detailed description of the between-link interactions, it's direct use for SO is not sufficient to design efficient algorithms. In other words, the use of a detailed simulation-based description of between-link interactions is not sufficient for the design of efficient algorithms. To achieve efficiency, the SO algorithm needs to also embed an *analytical* description of the between-link interactions (i.e., occurrence and impact of spillbacks).

  These insights then allow us to formulate an analytical network model that describes these between-links interactions while being sufficiently scalable for the efficient simulation-based optimization of large-scale congested networks (Section 4). In other words, we both: (i) simplify past macroscopic model formulations for increased scalability and efficiency, and (ii) tailor them for the optimization of congested networks by enhancing their description of between-link interactions.

  These insights highlight the need to formulate analytical macroscopic node models suitable for the optimization of large-scale congested networks. They emphasize the importance of developing analytical and differentiable formulations of the occurrence and impact of vehicular urban spillbacks, and of using these formulations for large-scale network optimization. These node models can be used both as part of stand-alone analytical macroscopic network models, as well as in combination with higher-resolution simulation-based traffic models for the design of efficient SO algorithms. More generally, based on these insights, we expect analytical node models to also enable the design of efficient data-driven signal control algorithms that rely mostly on queue-length measurements. In other words, by combining such measurements with an *analytical* description of between-link interactions, efficient and scalable data-driven algorithms can be designed.

- **Scalable analytical network model formulation.** To the best of our knowledge, the most scalable and efficient SO algorithm is that of Osorio and Chong (2015). A detailed description of how the proposed formulation compares to that of Osorio and Chong (2015) is given in Sections 2 and 4.1. Compared to the Osorio and Chong (2015) formulation, the proposed model has enhanced scalability, provides a more accurate analytical description of the between-link interactions under congested conditions, while remaining computationally efficient to evaluate. For a network with n lanes, the proposed network model is formulated as a system of n nonlinear convex differentiable equations.

- **Large-scale networks and high-dimensional SO problems.** The case study of Section 4 pushes the boundary of the scale of networks and SO problems that can be efficiently addressed. The signal control case study of Osorio and Chong (2015) considers, to the best of our knowledge, an optimization problem with the largest-scale high-resolution network model to date. It considers a microscopic network model with 800 lanes. The optimization problem controls a set of 17 intersections and 121 lanes, leading to a decision vector of dimension 99. Chong and Osorio (2016) consider a time-dependent problem that controls the same set of 17 intersections, leading to a decision vector of dimension 198. The case study of Section 4 considers a microscopic network model with 3691 lanes. The optimization problem controls a set of 96 intersections and 930 lanes, leading to a decision vector of dimension 259. The optimization problem considers a simulation-based non-convex objective function with convex analytical constraints. This is considered a high-dimensional and challenging problem in the field of SO.

- **Signal control.** This paper contributes to the field of signal control. It provides evidence that detailed *analytical* modeling of between-link interactions is critical for the control of urban networks with such intricate traffic patterns. Recent simulation-based signal control studies have mostly considered small-scale networks with a couple of arterials and with simple, mostly linear, network topologies (Park et al.; 2009; Stevanovic et al.; 2008; Yun and Park; 2006). This paper considers a large-scale microscopic model with over 3600 lanes arranged in an intricate grid topology. A total of 96 intersections are controlled. This is considered a high-dimensional network and intricate problem in the field of signal control.

- **State-of-practice.** This work is carried out in collaboration with the New York City Department of Transportation (NYCDOT). It presents two case studies of networks within New York City: a Queensboro Bridge network (Section 3) and a Midtown Manhattan network (Section 4). Current practice in the design of signal plans for these areas uses commercial software to design the signal plans. The software relies on simple traffic models. The derived signal plans are then embedded within these high-resolution simulators in order to provide a more detailed evaluation of their performance. Hence, the simulator is only used after the signal design process as an evaluation or validation tool. The results of this paper contribute to the state-of-practice by enabling the agency to systematically and efficiently use their high-resolution simulator within the signal design process. In other words, this work allows the agency to systematically use the simulation model within the signal optimization algorithm. This work complements NYCDOT's ongoing work in the signal control of Manhattan, such as their *Midtown in Motion* initiative (Xin et al.; 2013).

- **Manhattan case studies.** The case study of Section 4 controls a set of 96 intersections within an area of Manhattan. Other signal control strategies illustrated with Manhattan case studies include a simulation analysis for a network with 9 intersections (Spall and Chin; 1997), and both a simulation and an empirical analysis for a network with 7 intersections (Rathi; 1988). To the best of our knowledge, this paper considers the largest and most intricate Manhattan simulation-based optimization problem addressed so far.

Section 2 gives a brief description of the types of urban traffic simulators considered. It presents the main ideas of the SO methodology used in this paper. Section 3 identifies the analytical information provided by the macroscopic traffic models to the SO algorithms that is necessary for efficiency. Numerical results for a Queensboro Bridge case study are presented. The results of Section 3 are used in Section 4 to formulate a macroscopic model that is sufficiently scalable and efficient for high-dimensional SO problems of large-scale congested urban networks. The macroscopic model is embedded within an SO algorithm and used to address a Midtown Manhattan case study. The main conclusions of this paper and discussions of ongoing work are presented in Section 5. The SO algorithm is given in Appendix A.

2 Simulation-based optimization

This paper considers high-resolution simulation-based urban traffic models. These simulators are often stochastic models. They provide a detailed description of the underlying network supply (e.g., traffic management strategies). They describe demand at the scale of individual travelers or individual vehicles, and use disaggregate behavioral models to describe how travelers make pre-trip and en-route travel decisions. They account for the heterogeneity of traveler behavior. A given simulation run involves sampling a population of vehicles or travelers, each with their own set of travel decisions. For instance, in the case study of Section 4.2, one simulation run samples approximately 28,000 vehicles. The behavior of each vehicle is simulated by sampling from a variety of probabilistic behavioral models, such as route choice, car-following and lane-changing.

We use the general SO framework of Osorio and Bierlaire (2013). We summarize here its main ideas. The algorithm is given in Appendix A. For algorithmic details, we refer the reader to Osorio and Bierlaire (2013). The family of SO problems considered is formulated as follows.

$$\min_{x \in \Omega} f(G(x, z; p)) \tag{1}$$

where the purpose is to minimize a function f of a given stochastic performance measure G, x denotes the deterministic continuous decision vector, z denotes other endogenous simulation variables, and p denotes the exogenous simulation parameters. For example, in a signal control problem, G can represent link or network queue-lengths, f may represent the expectation operator (i.e., $f(G(x, z; p)) = E[G(x, z; p)]$), x can denote signal timing variables such as green times. The vector z can represent route choice decisions or signalized link flow capacities, while p accounts for network topology, lane attributes or exogenous prevailing traffic management strategies (e.g., lane-use priorities, pricing). The feasible region, Ω, consists of a set of general, typically nonconvex, deterministic, analytical and differentiable constraints. For instance, in signal control, the constraints can include lower bounds for green times.

The simulation-based objective function f is not known in closed-form. It can only be estimated via simulation. In order to obtain an accurate estimate, numerous simulation replications are needed. The use of high-resolution simulators leads to functions f that are typically nonconvex, and to computationally costly to evaluate replications. Hence, Problem (1) is difficult to address.

At every iteration k of the algorithm, a subproblem of the following form is solved.

$$\min_{x \in \Omega} m_k(x; \beta_k) \tag{2}$$

$$h(x, y; q) = 0. \tag{3}$$

Problem (1) differs from Problem (2)-(3) in two ways. First, the problem is constrained by an additional set of constraints (3). This constraint function h represents an analytical macroscopic traffic model with endogenous variables y (e.g., link densities) and exogenous parameters q (e.g., network topology). In this paper, it is formulated as an analytical differentiable system of nonlinear equations. Second, the simulation-based objective function f is replaced with an analytical function m_k. The latter is known as a metamodel. It is iteration-specific, i.e., at every iteration of the SO algorithm, a new analytical approximation of f is used. It depends on the decision vector x and on a vector of parameters β_k.

At every iteration k of the SO algorithm, the two main steps are: (i) the metamodel parameters β_k are fitted such as to minimize a distance metric between simulation observations (i.e., estimates of f) and m_k, (ii) Problem (2)-(3) is solved; its solution, known as the *trial point*, is then evaluated with the simulator. As the iterations advance, more points are simulated, which can increase the metamodel accuracy and lead to trial points with improved performance.

The function m_k is defined as:

$$m_k(x; \beta_k) = \beta_{k,0} f_A(x, y; q) + \phi(x; \beta_{k,1}, \ldots, \beta_{k,D}), \tag{4}$$

where:

- ϕ denotes a general-purpose analytical and differentiable function. It is specified as a quadratic polynomial in x with diagonal second-derivative matrix and with D coefficients $(\beta_{k,1}, \ldots, \beta_{k,D})$.

- f_A denotes a problem-specific analytical approximation of the simulation-based objective function f (Eq. (1)), it is derived by the analytical macroscopic traffic model h. The function f_A is scaled by the scalar coefficient $\beta_{k,0}$. The metamodel parameter vector is defined as $\beta_k = (\beta_{k,0}, \ldots, \beta_{k,D})$.

In the metamodel literature, ϕ is known as a functional model, while f_A is known as a physical model. The purpose of ϕ is to enable a general-purpose approximation of the objective function f, and the choice of a quadratic polynomial guarantees asymptotic convergence of the algorithm. The purpose of f_A is to provide a problem-specific analytical approximation of f. The metamodel m_k can be interpreted as an approximation, f_A, of f derived by an analytical macroscopic model, which is corrected parametrically by a scaling factor and an additive polynomial term, ϕ.

In summary, the SO framework relies on two main ideas. First, at every iteration k, we replace the unknown simulation-based objective function f with an analytical and differentiable function m_k. This allows the use of a variety of traditional optimization algorithms (e.g., gradient-based algorithms) to solve Problem (2)-(3). Second, f_A provides the algorithm with analytical problem-specific structural information. More specifically, it provides an analytical, differentiable and global approximation of the mapping between the decision vector x and the unknown objective function f. As a global approximation, it allows the algorithm to quickly identify subregions of the feasible region with good performance. This allows the algorithm to identify points with good performance even when few simulation observations are available. The key to designing a computationally efficient SO algorithm lies in f_A. By combining information from the simulator with problem-specific information from the analytical network model, the simulator is no longer used as a black-box, as in traditional SO algorithms (Stevanovic et al.; 2008; Park et al.; 2009).

Past signal control work has shown that the above metamodel (Eq. (4)) outperforms both: (i) a metamodel that consists only of the general-purpose component (i.e., the metamodel consists only of ϕ), and (ii) an approach that uses only the analytical traffic model, f_A (Osorio and Nanduri; 2015a; Osorio and Chong; 2015; Osorio and Nanduri; 2015b; Chen et al.; 2012; Osorio and Bierlaire; 2013). In other words, past studies have shown the added value for optimization of combining information from the simulation-based and the analytical traffic models.

The focus of this paper is on the design of SO algorithms for large-scale networks and high-dimensional problems. The main challenge in the use of this framework is the formulation of models h that provide a good approximation f_A of f, while also being: analytical, differentiable, scalable and efficient to evaluate. In particular, Problem (2)-(3), which is constrained by h, is solved at every iteration of the SO algorithm. Solving it efficiently requires the formulation of a highly efficient model h.

As the scale of the networks and the dimension of the SO problems increase, the computational efficiency of the SO algorithm is challenged because it is constrained by an analytical network model. Osorio and Chong (2015) recently proposed a macroscopic model h suitable for SO problems with large-scale networks. More specifically, for a network with n lanes, the model of Osorio and Chong (2015) has $\mathcal{O}(3n)$ complexity, i.e., the dimension of the system of equations represented by h (Eq. (3)) is $3n$. This formulation scales linearly with network size. Nonetheless, for the Midtown Manhattan case study of this paper (Section 4.2), the formulation of Osorio and Chong (2015) is not sufficiently scalable.

In this paper, we propose a formulation for h that achieves two goals. Compared to the Osorio and Chong (2015) formulation, the proposed model: (i) has enhanced scalability: it is formulated as a system of n convex differentiable equations, (ii) remains computationally efficient to evaluate, and (iii) provides a more accurate analytical description of the between-link interactions under congested conditions. The latter is shown in Section 3 to be the key structural information provided to the SO algorithm by the analytical traffic models for the control of congested urban networks.

3 Analytical structural information for congested network optimization

This section investigates what structural analytical information provided by the model h is needed for the design of efficient SO algorithms for congested urban networks. This understanding is then used in Section 4 to design an SO algorithm with enhanced efficiency and scalability. To investigate this, we consider two SO algorithms that differ only in the analytical information they provide to the algorithm. More specifically, they differ only in the choice of the analytical model used as part of the metamodel (h of Eq. (3)). The first uses an analytical traffic model that accounts for the occurrence and impact of vehicular spillbacks, while the second does not account for these between-link interactions. Both use the SO framework of Osorio and Bierlaire (2013).

Section 3.1 presents the two analytical traffic models. Both are used to address a simulation-based signal control problem for the Queensboro Bridge (QBB) network. Section 3.2 presents the QBB network. Section 3.3 formulates the optimization problem. Section 3.4 presents the case study results.

3.1 Analytical models

Both of the analytical traffic models are based on a combination of traffic flow theory ideas and queueing network theory ideas. The first model is formulated for general queueing networks in Osorio and Bierlaire (2009). Its formulation for urban road networks is given in Osorio (2010, Chap. 4). Each lane of a road network is modeled as a finite space capacity queue. In queueing theory spillbacks are referred to as blockings. The model uses the notion of *blocking* to describe the occurrence and effects of vehicular spillbacks. Link, rather than network, models of vehicular traffic based on queueing theory include Osorio and Flötteröd (2015); Osorio et al. (2011); Heidemann (2001); Jain and Smith (1997); Heidemann (1996, 1994); Tanner (1962). The considered model is formulated as follows. The index i refers to a given queue.

γ_i	external arrival rate;
$\hat{\lambda}_i$	effective arrival rate;
μ_i	service rate;
$\hat{\mu}_i$	effective service rate;
ρ_i	traffic intensity;

ℓ_i space capacity;

N_i queue-length;

$P(N_i = \ell_i)$ probability of queue i being full;

p_{ij} routing probability from queue i to queue j;

$\mathcal{D}_i$ set of downstream queues of queue i.

$$\begin{cases} \hat{\lambda}_i = \gamma_i(1 - P(N_i = \ell_i)) + \sum_j p_{ji}\hat{\lambda}_j & \text{(5a)} \\ P(N_i = \ell_i) = \dfrac{1-\rho_i}{1-\rho_i^{\ell_i+1}}\rho_i^{\ell_i} & \text{(5b)} \\ \rho_i = \dfrac{\hat{\lambda}_i}{(1 - P(N_i = \ell_i))\hat{\mu}_i} & \text{(5c)} \\ \dfrac{1}{\hat{\mu}_i} = \dfrac{1}{\mu_i} + \left(\sum_j p_{ij} P(N_j = \ell_j)\right)\left(\sum_{j \in \mathcal{D}_i} \dfrac{\hat{\lambda}_j}{\hat{\lambda}_i \hat{\mu}_j}\right). & \text{(5d)} \end{cases}$$

Equation (5a) is a flow conservation equation that defines the expected flow of a given queue i as the sum of the expected flow arising from outside the network (this is represented by γ_i and can be interpreted as the expected number of trips that start at queue i) and of the expected flow arising from upstream links (this is given by the summation term). Equation (5b) defines the probability that a queue is full. This probability is also known as the spillback probability or the blocking probability. It represents the probability that the underlying lane spills back. This expression is obtained by modeling the underlying lane as an M/M/1/ℓ queue (e.g., Bocharov et al.; 2004). This probability is a function of the lane's space capacity ℓ_i and the lane's traffic intensity ρ_i. The latter is defined by Equation (5c), and is defined as the ratio of expected demand to expected supply. Note that the use of finite space capacity queues (i.e., assuming $\ell_i < \infty$) allows for any non-negative value of ρ_i. In particular, values larger than one are allowed. This means that highly congested scenarios where expected demand exceeds expected supply can be accounted for. This is particularly important when considering peak period scenarios, as in this paper.

Equation (5d) describes the effective service rate of queue i, denoted $\hat{\mu}_i$. This notion is similar to an effective flow capacity term. The term *effective* relates to the fact that the flow capacity of a link can be affected by downstream traffic conditions (e.g., spillback of a downstream queue). The inverse of the effective service rate (i.e., $1/\hat{\mu}_i$) is known as the expected effective service time. It is defined as the sum of: (i) the expected service time (term $1/\mu_i$), where μ_i is a flow capacity variable that depends only on attributes of the underlying lane (e.g., maximum speed), i.e., it does not depend on any downstream traffic conditions, and of (ii) the expected blocked time. The latter represents the expected additional time a vehicle spends in queue i while waiting for a spillback to dissipate at a downstream queue. The expected blocked time is approximated by the product of: (i) the probability that a downstream link spills back (summation in the first parenthesis), and of (ii) the expected unblocking time (summation in the second parenthesis). The expected unblocking time represents the expected time it takes for a spillback at a downstream queue to dissipate. It is Equation (5d) which provides a differentiable description of between-link interactions under congested conditions. In other words, it describes the impact on the flow capacity of queue i due to spillbacks from its downstream queues. It is through this equation that the model provides a description of the across-node spatial propagation of congestion.

The second model differs from the first in that it does not account for the between-link (i.e., between-queue) spillback interactions. It is formulated as follows.

$$\begin{cases} \hat{\lambda}_i = \gamma_i(1 - P(N_i = \ell_i)) + \sum_j p_{ji}\hat{\lambda}_j & \text{(6a)} \\ P(N_i = \ell_i) = \dfrac{1-\rho_i}{1-\rho_i^{\ell_i+1}}\rho_i^{\ell_i} & \text{(6b)} \\ \rho_i = \dfrac{\hat{\lambda}_i}{(1 - P(N_i = \ell_i))\mu_i}. & \text{(6c)} \end{cases}$$

Equations (6a) and (6b) are identical to (5a) and (5b), respectively. Equation (6c) differs from (5c) in that the expected supply is represented by μ rather than by $\hat{\mu}$. Recall that $\hat{\mu}$ is defined as the effective service rate. It differs from μ in that is accounts for the impact of downstream spillbacks on the links flow capacity. If for a given queue, the spillback probability of all of its downstream queues is zero, then $\hat{\mu}$ equals μ. Otherwise, $\hat{\mu} < \mu$. In other words, the model described by the System of Equations (6) accounts for the within-link congestion propagation (through Eq. (6b)), yet does not account for the between-link propagation.

To summarize, the two SO methods that we compare differ only in their analytical macroscopic model. The first (Eq. (5)) is referred to as the *SO-Spill* approach. The second (Eq. (6)) is referred to as the *SO-No-Spill* approach. Table 1 summarizes which traffic models are used by each of the two compared approaches.

3.2 Queensboro Bridge network

We consider a network within New York City referred to as the Queensboro Bridge (QBB) network. Figure 1 displays a map of central Manhattan, the QBB area of interest is delimited by the oval. Figure 2(a) displays a detailed map of the QBB area, the corresponding network model is displayed in Figure 2(b).

	Microscopic (simulation-based)	Macroscopic (analytical)	
		Eq. (5)	Eq. (6)
SO-Spill	√	√	
SO-No-Spill	√		√

Table 1: Traffic models used by the compared methods.

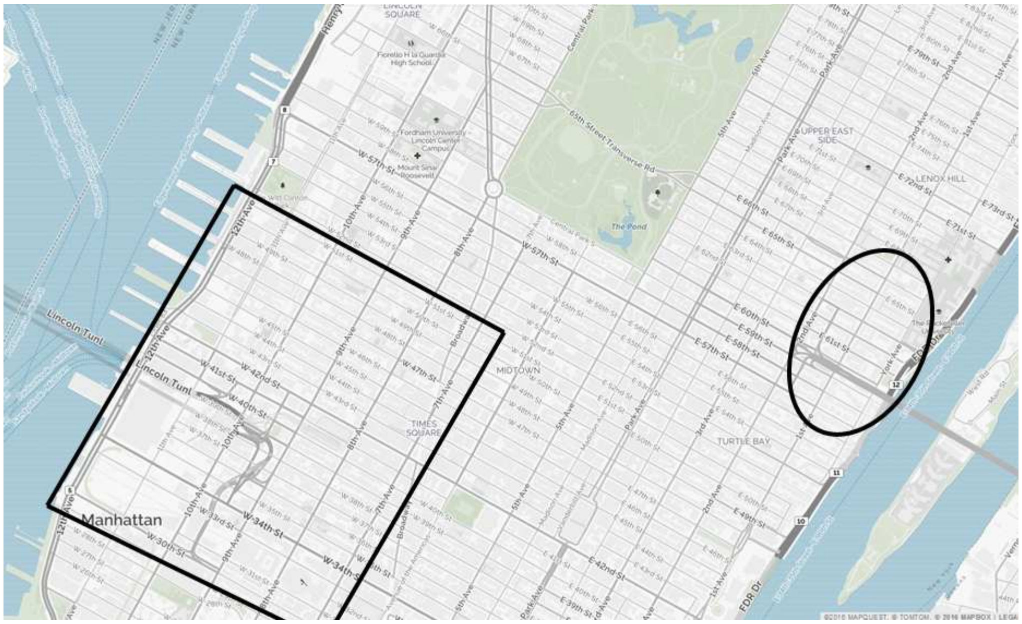
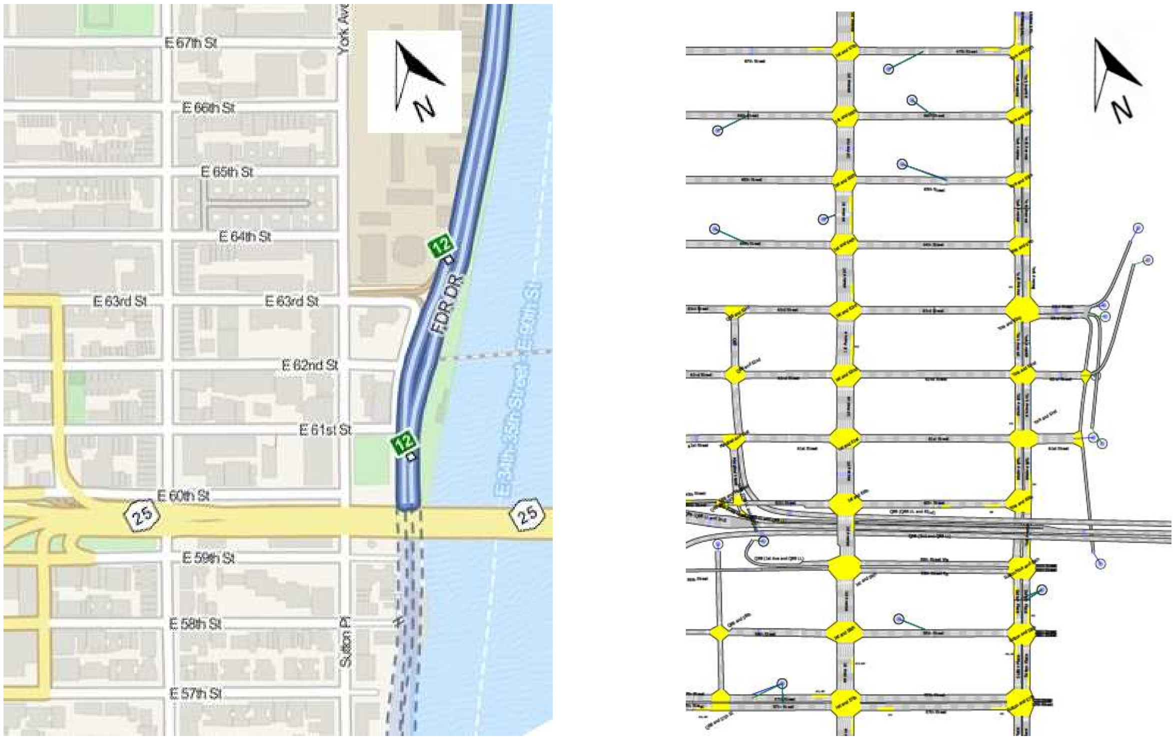


Figure 1: Central Manhattan with networks of interest delimited by an oval (Queensboro Bridge network) and a rectangle (Midtown Manhattan network) (MapQuest.com, Inc; 2015).



(a)

(b)

Figure 2: Queensboro bridge network of interest: map (left plot) (MapQuest.com, Inc; 2015) and model (right plot).

The design of suitable signal plans, and more generally of suitable traffic management strategies, for this area is a major challenge due to the intricate traffic patterns that are caused by the factors mentioned in the first paragraph of Section 1 (e.g., short links, grid topology) and by the following additional factors.

The network is highly congested with morning peak period vehicular traffic, which is the focus of this paper, in the order of 12,000 trips per hour. Congestion is homogenously distributed, i.e., it arises in all directions of travel. Additionally, all cross-streets are considered important. Hence, the congestion patterns do not reveal a clear hierarchy between intersections or between roads. The network has intricate multi-modal traffic patterns (e.g., cars, trucks, buses, bikes and pedestrians). It has considerable pedestrian traffic. For instance, the pedestrian volume at the 1st Avenue intersection with 59^{th} Street can be of the order of 500 pedestrians per hour during morning peak period (NYCDEP; 2006). Hence, the signal plans at every intersection are constrained by significant pedestrian crossing times.

This area includes the Ed Koch Queensboro Bridge, which connects Queens to Manhattan. This is the busiest NYCDOT bridge, with average weekday traffic in the order of 178,000 vehicles (NYCDOT; 2015). During the morning peak hour 8-9am, which is the focus of this paper, an estimated 5,770 vehicles cross the bridge in the Manhattan-bound direction (NYCDOT; 2012). The Queensboro Bridge also serves as a key transit corridor that carries approximately 694 buses and 16,000 bus passengers on a typical weekday. It is the busiest express bus route exiting Manhattan, connecting neighborhoods in Midtown Manhattan and eastern Brooklyn and Queens (NYCDOT; 2011). During the peak period, streets adjacent to the bridge are typically congested, and travel times along the bridge increase significantly (NYCDOT; 2010). The occurrence of congestion along access and egress links of the bridge can have significant impacts on congestion of links in both Manhattan and Queens, and on transit service. Vehicular queues during peak periods often require traffic enforcement agents to prevent spillback and gridlock. This further motivates the design of signal plans that mitigate the spatial propagation of congestion. The importance of this bridge for both Manhattan and Queens traffic indicates that even small local changes to traffic operations in this area can trigger significant larger-scale impacts. The study area also includes several exit and entrance links to the Franklin D. Roosevelt East River Drive (FDR Drive), which is an important limited-access facility that allows traffic to travel along Manhattan's east edge. Local streets near the FDR Drive (e.g. York Avenue at 63rd Street) often experience heavy turn volumes and high levels of congestion during peak hour.

The network model (Fig. 2(b)) consists of 134 roads, 313 lanes, 27 signalized intersections and 5 non-signalized intersections. The signal control problem, formulated in Section 3.3, determines the signals of 26 signalized intersections. This leads to a total of 64 endogenous signal phases (i.e., the decision vector is of dimension 64) that control the flow of a total of 120 queues. We consider the 8-9am morning peak period. The expected demand, represented by a static origin-destination matrix, consists of over 11,500 car trips and over 750 truck trips that are distributed across a set of 55 origin-destination pairs. The calibrated microscopic simulation model is developed with the Aimsun microscopic simulation software (TSS; 2013).

3.3 Signal control problem

Reviews of traffic signal control terminology are given in Lin (2011); Osorio (2010, Appendix A); Papageorgiou et al. (2003). The current approach to signal control in the QBB area for peak period traffic is the use of fixed-time signal plans, also known as time-of-day or pre-timed plans. They are periodic plans with a period or cycle time of typically 90 seconds or 120 seconds for each intersection. They are designed offline for a given time period (e.g., morning peak period) based on model forecasts and historical traffic patterns. Fixed-time signal plans are the most traditional form of signal timing. Unlike adaptive or traffic-responsive plans, they do not vary with real-time changes in demand or supply. They are the standard practice in cities with sparse real-time traffic data. Major cities with abundant real-time traffic data, such as New York, use fixed-time plans for time periods with high and uniformly distributed congestion levels. Fixed-time plans are also used to design traffic responsive plans, where an offline determined fixed-time plan is selected in real-time from a portfolio of plans, as in Chen et al. (2015).

The decision variables of our problem are known as the green splits of the signal phases, i.e., the ratio of green times to cycle times of the phases. All other traditional signal plan variables (e.g., cycle times, offsets, stage structure) are assumed fixed. We consider a problem where the signals of all endogenously controlled intersections are determined jointly. In order to formulate the signal control problem, we introduce the following notation:

c_i	cycle time of intersection i;
d_i	fixed cycle time of intersection i;
$x(j)$	green split of phase j;
x_L	vector of minimal green splits;
N_l	number of vehicles on link l;
$\mathscr{I}$	set of intersection indices;
$\mathcal{L}$	set of link indices;
$\mathscr{P}_I(i)$	set of phase indices of intersection i.

The optimization problem is formulated as follows.

$$\min_x f(x) = \sum_{l \in \mathcal{L}} E[N_l(x, z; p)] \tag{7}$$

$$\sum_{j \in \mathscr{P}_I(i)} x(j) = \frac{c_i - d_i}{c_i}, \ \forall i \in \mathscr{I} \tag{8}$$

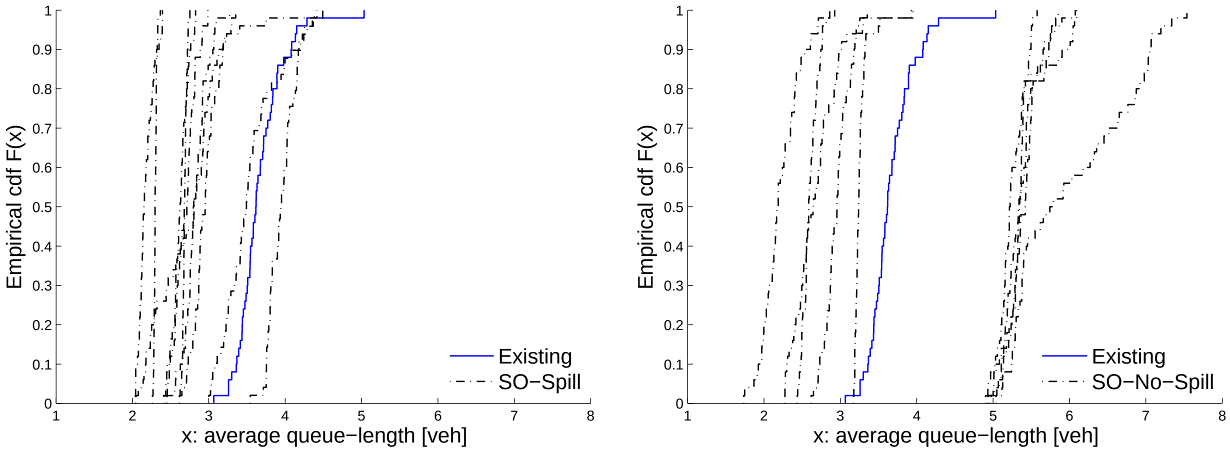
$$x \geq x_L, \tag{9}$$

where x is the decision vector, $f(x)$ represents the (unknown) simulation-based objective function defined as the expected number of vehicles in the network, i.e., it is the sum over all links of the expected number of vehicles on each link. The summation is over all links in the network, both controlled and non-controlled. Intersection i has a fixed (i.e., exogenous) cycle time, c_i. Certain signal phases within the cycle (e.g., all-red phases) are considered fixed. The sum of the durations of these fixed phases is referred to as the fixed cycle time, d_i. Hence, the left-hand side of Constraint (8) denotes the proportion of the cycle time that is endogenous (i.e., the proportion that can be allocated to the endogenous signal phases). Constraint (8) ensures that for each intersection the endogenous green times sum to the total available cycle time. Lower bounds are used to ensure a minimal duration for all green phases (Constraint (9)).

3.4 Numerical analysis

This section compares the performance of the two SO algorithms: SO-Spill and SO-No-Spill. One run of a given SO algorithm involves calling the stochastic microscopic simulator. The outputs of an SO run are therefore stochastic. Thus, in order to evaluate the performance of a given algorithm we run it 10 times with the same initial point. We initialize each run with the existing NYCDOT signal plan as the initial point. For a given algorithmic run, we allow for a total of 150 simulation calls, i.e., the simulation budget is set to 150. Each SO run yields as output a *proposed* signal plan. In summary, we run each SO algorithm 10 times to obtain 10 proposed signal plans. In order to evaluate the performance of a proposed signal plan, we run 50 simulation replications. For each replication, we estimate the objective function. We construct a cumulative distribution function (cdf) based on these 50 objective function estimates. We then compare the distributions obtained from the different signal plans.

Figure 3(a) displays 11 curves. Each curve considers a given signal plan. Each curve displays the cdf of the average queue-length of a given signal plan. The 10 dashed cdf's correspond to signal plans proposed by SO-Spill. The solid blue cdf corresponds to the existing NYCDOT signal plan. As described above, each curve is built based on observations from 50 independent simulation



(a)

(b)

Figure 3: Comparison of the average queue-lengths for the signal plans proposed by SO-Spill (3(a)) and by SO-No-Spill (3(b)).

replications. The x-axis is the average queue-length. This average is obtained as an average across all links of the network of the average (across time) link queue-length. If we multiply the average queue-length by the total number of links in the network, then we obtain an estimate of the objective function. The y-axis is the estimated cumulative probability. For example, consider a curve that passes through a given point $(x, y) = (3, 0.9)$. It can be interpreted as: for that signal plan 90% (i.e., a fraction of 0.9) of the observations have values smaller than 3. In other words, of the 50 simulation observations of the average queue-lengths, 90% have values smaller than 3. Thus, for a given signal plan, the more its cdf is shifted to the left, the higher the proportion of simulation replications with low average queue-length observations, i.e., the better its performance. In Figure 3(a), 9 out of the 10 signal plans proposed by SO-Spill lead to smaller average queue-lengths than that of the existing signal plan. One plan performs worse than the existing plan.

Figure 3(b) also displays 11 curves: the 10 black dashed cdf curves correspond to the signal plans proposed by SO-No-Spill, the solid blue cdf corresponds to the existing NYCDOT signal plan. Figure 3(b) indicates that 5 out of the 10 plans proposed by SO-No-Spill have better performance than the existing plan, the remaining 5 plans have significantly worse performance than the existing plan. Figures 3(a) and 3(b) have the same x-axis range, and can be directly compared.

We analyze the ability of the SO-Spill method to derive signal plans that mitigate the spatial propagation of congestion. Based on criteria provided by NYCDOT, a proposed signal plan should reduce the spatial propagation of congestion, yet should not deteriorate the network throughput.

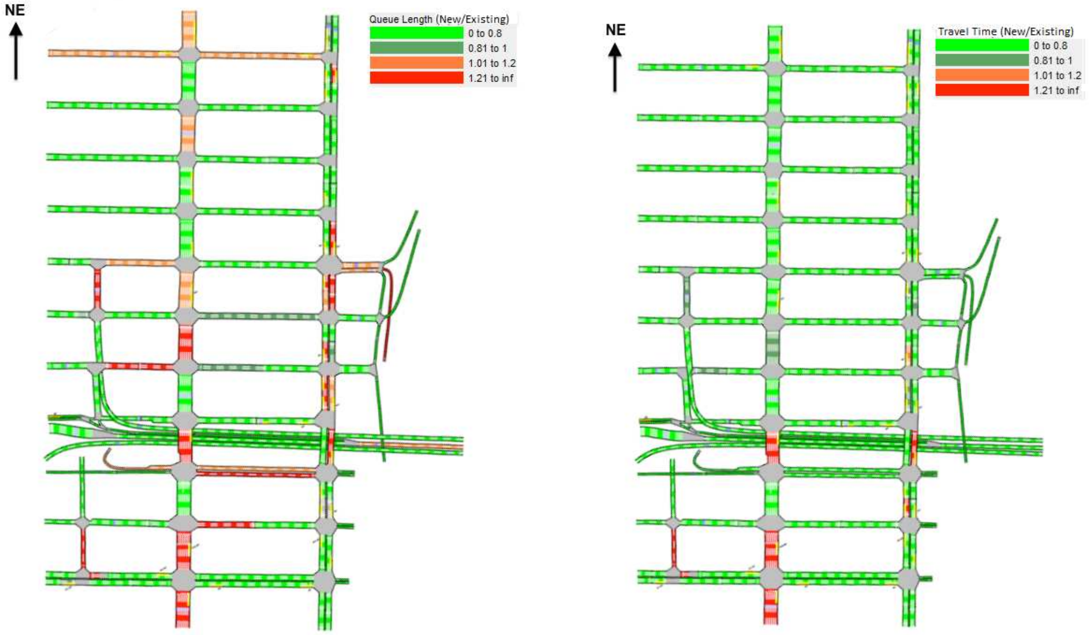
Hence, among the plans proposed by the SO-Spill algorithm, we consider as *best* plan that with the lowest average queue-length and with a network throughput that is statistically equivalent or better than that of the existing NYCDOT plan. Figure 4(a) compares the performance of the *best* SO-Spill plan and of the existing NYCDOT plan. For a given signal plan (existing or SO-Spill), we estimate the average link queue-length. This average is estimated as an average over both the simulation period (8-9am) and over the 50 simulation replications. The links of Figure 4(a) are colored according to the ratio of average link queue-length of the SO-Spill plan and the existing plan. Links are colored as follows: a decrease of average queue-length (of the SO-Spill plan compared to the existing plan) of more than 20% corresponds to light green, a decrease within 0 and 20% corresponds to dark green, an increase between 0 and 20% corresponds to orange and an increase of more than 20% corresponds to red. The majority of the links, and in particular almost all cross-streets, have significantly reduced queue-lengths. The SO-Spill plan leads to an improvement in average queue-lengths at the network level (i.e., improvement in the objective function as shown in Figure 3(a)) as well as an improvement at the link-level (as shown in Figure 4(a)). Hence, the SO-Spill plan achieves the goal of mitigating the spatial propagation of congestion.

Figure 4(b) is constructed just as Figure 4(a) but considers the average link travel time. It shows that under the SO-Spill plan almost all links have an improvement in average link travel time of more than 20% compared to the existing signal plan.

The main results of this section are two-fold. First, they show that the SO-Spill algorithm outperforms the SO-No-Spill algorithm. This indicates that there is a significant added value in providing the SO algorithm with an analytical description of the between-link interactions, i.e., of the occurrence of spillbacks and their impact on upstream link performance. Second, they show that the SO-Spill algorithm identifies signal plans that lead to significant improvements in performance compared to the existing plan, both at the network level and at the link level. They achieve the goal of limiting the spatial propagation of congestion within the intricate QBB network. For a more detailed analysis of the performance of the algorithms, as well as implementation details, we refer the reader to Osorio et al. (2015).

4 Analytical scalable network model

The analysis of Section 3 indicates the importance of providing SO algorithms with analytical information of between-link interactions. This section builds upon these results to formulate an analytical model that accounts for between-link interactions while also being suitable for large-scale network simulation-based optimization (Section 4.1). For a network with n lanes, the proposed model con-



(a) Ratio of average link queue-length

(b) Ratio of average link travel time

Figure 4: Comparison of the spatial performance of the SO-Spill plan and the existing plan.

sists of a system of n nonlinear convex equations. This new formulation is sufficiently scalable and computationally efficient to address the signal control problem for a large-scale network in Midtown Manhattan. The network is presented in Section 4.2. The numerical results are presented in Section 4.3.

4.1 Model formulation

In the network model defined by the System of Equations (5), it is Equation (5d) that accounts for the interactions between links under congested conditions. Recall from Section 3.1 that this equation describes the impact of spillback on the flow capacity of upstream links. More specifically, Equation (5d) approximates the expected effective service time of queue i (represented by $1/\hat{\mu}_i$). The latter is defined as:

$$\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + E[B_i], \tag{10}$$

where $E[B_i]$ denotes the expected blocked time at queue i. Equation (10) states that the expected time that a vehicle in queue i undergoes service is defined as the expected time it undergoes service without the occurrence of blocking (represented by $1/\mu_i$) and the expected time it is blocked (represented by $E[B_i]$). The latter is defined as:

$$E[B_i] = \sum_{j \in \mathscr{D}_i} p_{ij} P(N_j = \ell_j) E[B_{ij}], \tag{11}$$

where $E[B_{ij}]$ denotes the expected blocked time at queue i due to blocking (i.e., spillback) from downstream queue j. Equation (11) states that the expected blocked time at queue i due to spillback on queue j is defined by: the probability that a vehicle in queue i has chosen queue j as its downstream destination queue (represented by the turning or routing probability p_{ij}), the probability that this destination queue is full (represented by $P(N_j = \ell_j)$) and the expected time until the vehicular spillback from this blocked destination queue j dissipates and allocates an available space for the vehicle in queue i (represented by $E[B_{ij}]$).

Equation (5d) was obtained by approximating the expected blocked time as follows:

$$\begin{cases} E[B_i] = \left(\sum_{j \in \mathscr{D}_i} p_{ij} P(N_j = \ell_j) \right) \left(\sum_{j \in \mathscr{D}_i} E[B_{ij}] \right) & \text{(12a)} \\ E[B_{ij}] = \dfrac{\hat{\lambda}_j}{\hat{\lambda}_i \hat{\mu}_j}. & \text{(12b)} \end{cases}$$

The above approximations are derived and discussed in detail in Osorio and Bierlaire (2009, Section 4.2.3 and Appendix).

The model formulated in this paper considers the expression of the expected blocked time (Eq. (11)), i.e., we do not carry out the approximation of Equation (12a). We insert Equation (12b) into (11) to obtain:

$$E[B_i] = \sum_{j \in \mathscr{D}_i} p_{ij} P(N_j = \ell_j) \frac{\hat{\lambda}_j}{\hat{\lambda}_i \hat{\mu}_j}. \tag{13}$$

Inserting Equation (13) into Equation (10), we obtain:

$$\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + \sum_j p_{ij} P(N_j = \ell_j) \frac{\hat{\lambda}_j}{\hat{\lambda}_i \hat{\mu}_j}. \tag{14}$$

We then proceed as in Osorio and Chong (2015), i.e., we approximate the traffic intensity ρ_i with the *effective traffic intensity*, which is denoted $\hat{\rho}_i$ and is defined as the ratio of the effective arrival rate and the effective service rate (i.e., $\hat{\rho}_i = \hat{\lambda}_i / \hat{\mu}_i$). We then multiply Equation (14) with $\hat{\lambda}_i$ to obtain:

$$\hat{\rho}_i = \frac{\hat{\lambda}_i}{\mu_i} + \sum_j p_{ij} P(N_j = \ell_j) \hat{\rho}_j. \tag{15}$$

This gives the following network model formulation:

$$\begin{cases} \hat{\lambda}_i = \gamma_i(1 - P(N_i = \ell_i)) + \sum_j p_{ji}\hat{\lambda}_j & \text{(16a)} \\ P(N_i = \ell_i) = \dfrac{1-\hat{\rho}_i}{1-\hat{\rho}_i^{\ell_i+1}}\hat{\rho}_i^{\ell_i} & \text{(16b)} \\ \hat{\rho}_i = \dfrac{\hat{\lambda}_i}{\mu_i} + \sum_j p_{ij} P(N_j = \ell_j)\hat{\rho}_j. & \text{(16c)} \end{cases}$$

System (16) differs from the formulation of Osorio and Chong (2015) in that it is based on a more accurate approximation of the expected blocked time. Nonetheless, they both have similar scalability. They are both defined as a system of n linear, n quadratic and n non-quadratic convex equations.

In order to derive a more scalable model, we propose a formulation that consists of n convex equations. To achieve this, we consider arrival rates to be exogenous. More specifically, we solve System (16) once prior to optimization. This yields a vector of exogenous arrival rates, denoted $\tilde{\lambda}$. These are considered fixed throughout the optimization process. This leads to the following formulation:

$$\begin{cases} \hat{\rho}_i = \dfrac{\tilde{\lambda}_i}{\mu_i} + \sum_{j \in \mathscr{D}_i} p_{ij} P(N_j = \ell_j)\hat{\rho}_j & \text{(17a)} \\ P(N_i = \ell_i) = \dfrac{1-\hat{\rho}_i}{1-\hat{\rho}_i^{\ell_i+1}}\hat{\rho}_i^{\ell_i}, & \text{(17b)} \end{cases}$$

where $\hat{\rho}_i$ now denotes the ratio of exogenous expected demand $\tilde{\lambda}_i$ to expected supply $\hat{\mu}_i$. We insert Equation (17b) into (17a) to obtain:

$$\hat{\rho}_i = \frac{\tilde{\lambda}_i}{\mu_i} + \sum_{j \in \mathscr{D}_i} p_{ij} \left(\frac{1-\hat{\rho}_j}{1-\hat{\rho}_j^{\ell_j+1}}\hat{\rho}_j^{\ell_j} \right) \hat{\rho}_j. \tag{18}$$

In summary, for a network of n queues, the analytical traffic model consists of a system of n equations (18), which account for the impact of blocking and hence capture between-queue dependency information under congested conditions.

4.2 Midtown Manhattan case study

We consider an area within Midtown Manhattan. It is delimited in Figure 1 by a black rectangle. The simulation model considers 924 roads, 3691 lanes and 444 nodes. It is displayed in Figure 5. It

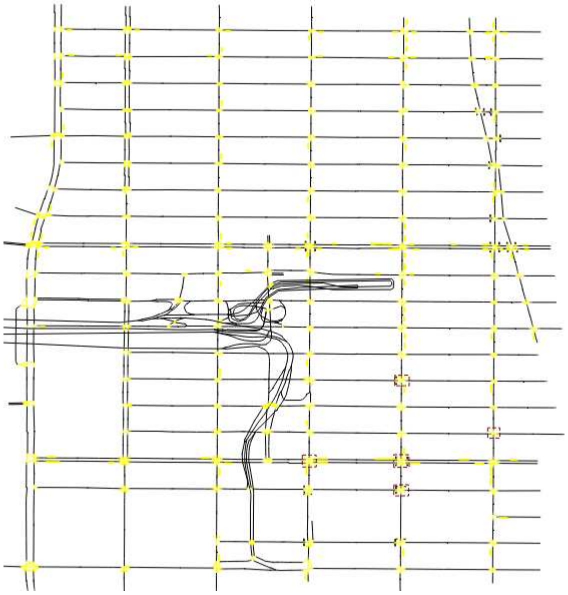


Figure 5: Midtown Manhattan network model

considers 101 controlled intersections. In this case study, we determine the signals of 96 controlled intersections, which control 930 lanes, and are defined as a total of 259 endogenous signal phases. In other words, the dimension of the decision vector is 259. We simulate traffic for the 3-6pm period, and determine signal plans for the 5-6pm evening peak hour. For this hour, the expected demand is over 27,000 car trips per hour, over 1,300 truck trips per hour and is distributed across over 2,600 OD pairs.

The optimization problem is constrained by Equations (8)-(9). The simulation-based objective function f is defined as the expected trip travel time. This function is equivalent to that of Equation (7) under the assumption of fixed network demand. The use of expected trip travel time as the objective function, instead of expected link density, allows for a more computationally efficient estimation of the objective function (since link-level statistics are no longer stored).

4.3 Numerical analysis

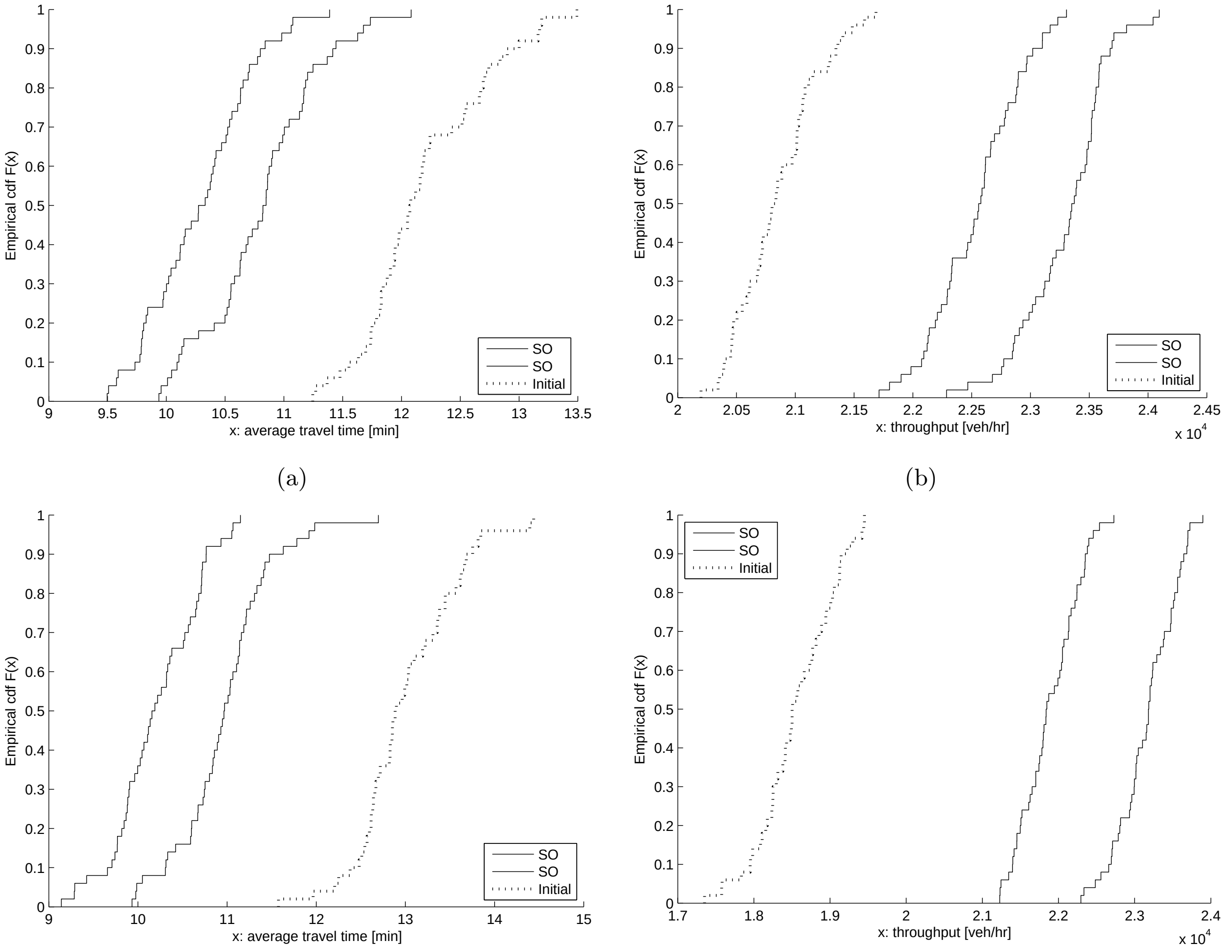
We use the model defined by (18) as the analytical network model of the SO algorithm, i.e., (18) represents the h function of Eq. (3). For each algorithmic run, we allow for a maximum of 50 simulation runs. In other words, the computational budget is set to 50. We consider two different initial points (i.e., signal plans) that are uniformly and randomly drawn from the feasible space defined by Equations (8) and (9). To sample uniformly, we use the code of Stafford (2006). For each initial point, we run the SO algorithm two times, each time allowing for 50 simulation runs. Thus, for each initial point, we derive two *proposed* signal plans. After optimization, we use the simulator to evaluate in detail the performance of the proposed and the initial signal plans. We consider two performance metrics: the average trip travel time, which is the objective function of our problem, and the network throughput, which is a metric that is of particular interest to the agency. In order to evaluate the performance of a signal plan, we proceed as in Section 3.4: we run 50 simulation replications and compare the cumulative distribution functions (cdf's).

Each row of plots of Figure 6 considers one random initial point. The left (resp. right) column plots evaluate the signal plans in terms of average trip travel (resp. network throughput). Each plot displays three curves. Each curve corresponds to one signal plan. The two solid curves correspond to the plans proposed by each of the two SO algorithm runs. The dotted curve corresponds to the initial plan. Each curve is a cdf of the performance measure (trip travel time or network throughput). The more the trip travel time curves are shifted to the left, the higher the proportion of simulation replications (of the 50 simulation replications) with low average trip travel time estimates, i.e., the better the performance of the signal plan. For the network throughput metrics, since higher values of throughput are desirable, the more the curves are shifted to the right, the better the performance of the signal plan.

Figure 6(a) shows that both plans identified by the algorithm lead to significantly better average trip travel time estimates than the initial plan. Figure 6(b) considers the same signal plans as in Figure 6(a), and compares them in term of their throughput. This figure shows that both of the proposed signal plans lead to significantly higher throughput values compared to the initial plan. The same conclusions hold for the signal plans derived with the second random initial plan (Figures 6(c) and 6(d)).

We provide a spatial comparison of the performance of a proposed plan with the existing NYCDOT plan. We consider the plans of Figure 6 and choose that with the best performance. This corresponds to the signal plan with the cdf that is the most to the right in Figure 6(a) and most to the left in Figure 6(b). Hereafter, we refer to this plan as the proposed plan. Figures 7 and 8 consider the existing signal plan and the proposed plan. For each plan, 50 simulation replications are carried out, and link-level statistics are collected. Figure 7 colors each link of the network according to the ratio of the average link travel time under the proposed plan and that under the existing plan. Links are colored green if there is an improvement (i.e., a ratio smaller than 1), and in yellow, orange and red tones if there is a deterioration (i.e., a ratio larger than 1). This figure indicates, that at the link-level there is a majority of links that have improved average travel times.

Figure 8 compares the signal plans according to the average link density. The majority of the



(a)

(b)

(c)

(d)

Figure 6: Cdf's of the average trip travel times (6(a) and 6(c)) and the network throughput (6(b) and 6(d)) considering two random initial points.

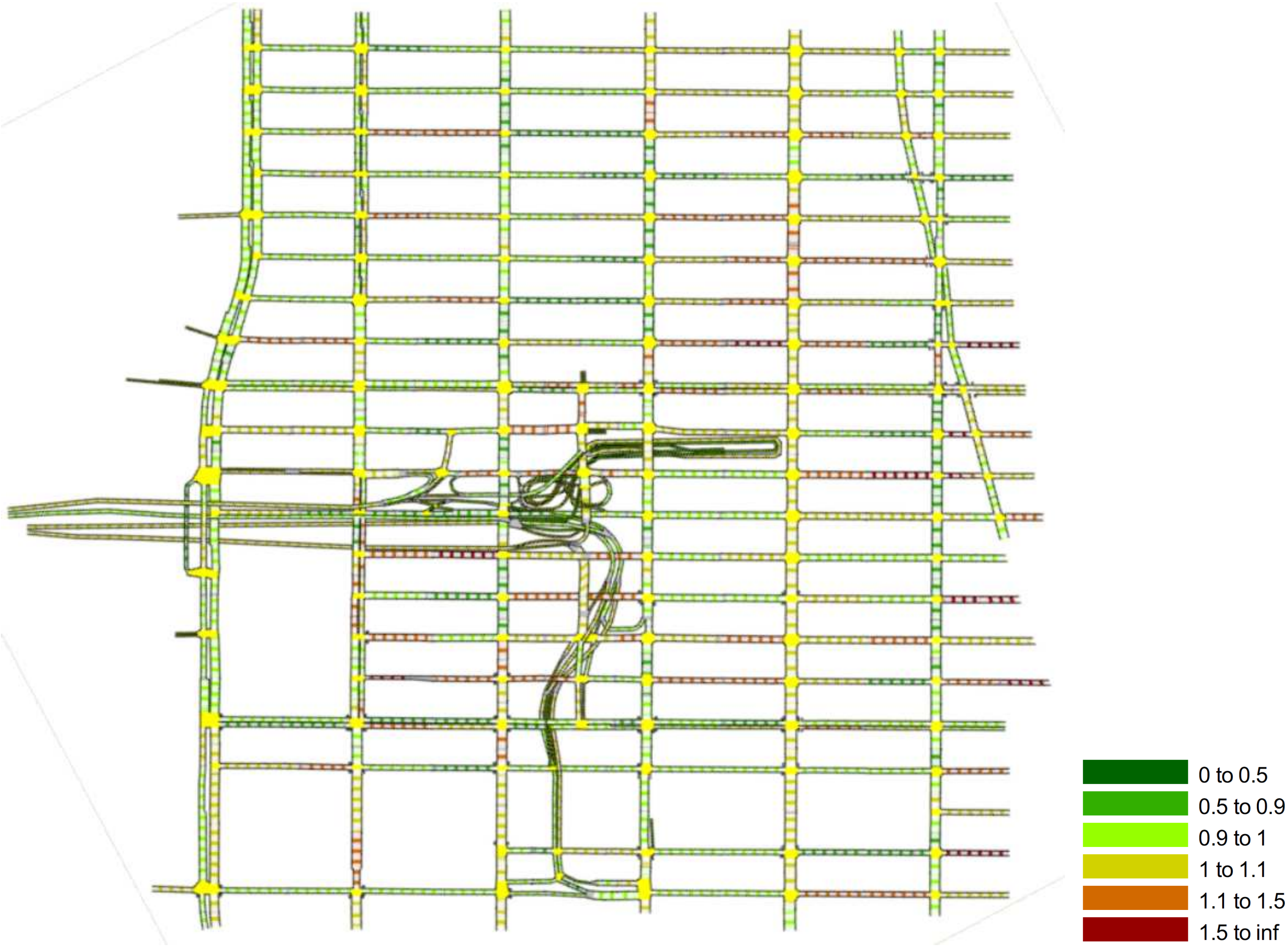


Figure 7: Ratio of the average link travel time under the proposed plan and that under the existing plan

links have lower average density values. The problem addressed considers all links to have equal importance or priority. Given the intricate traffic dynamics in this network and, in particular, the interactions and priorities between cross-streets and avenues, it would be of interest to formulate and address problems where travel directions are accounted for.

5 Conclusions

This paper considers a family of metamodel SO algorithms, where the metamodels combine information from the simulator with information from an analytical traffic model. Section 3 investigates what information provided by the analytical traffic model is necessary for the design of efficient SO algorithms for large-scale congested networks. The Queensboro Bridge case study indicates that providing the algorithm with an *analytical* description of the occurrence and impact of spillbacks is necessary for efficiency. The performance of a proposed signal plan is compared to that of an existing NYCDOT plan. The proposed plan improves performance both at the network level and at the link level. It mitigates the spatial propagation of congestion within the intricate QBB network. Section 4 formulates an analytical model with enhanced scalability and efficiency, suitable for the simulation-based optimization of congested urban networks. The Midtown Manhattan case study illustrates the ability of the proposed model to efficiently address a high-dimensional SO problem for a large-scale network with congested and intricate traffic patterns. The SO algorithm identifies, under tight computational budgets, signal plans that outperform the initial points. The performance of a proposed signal plan is also compared to that of an existing NYCDOT plan. It is shown to improve the average link travel times and average link densities, for the majority of links.

The results of this paper indicate that for congested urban networks with intricate traffic patterns there is a significant added value in providing the SO algorithms with analytical between-queue interaction information. The formulation of analytical and differentiable traffic models that describe between-queue interactions is of interest to address a variety of network design and operations problems. In particular, the findings of this paper contribute to inform the design of the next-generation of SO algorithms. For instance, analytical between-queue dependency can be used to improve sampling strategies (e.g., sampling of model improvement points as defined in Osorio and Bierlaire (2013)) or to develop more efficient statistical comparisons of point performance (e.g., ranking and selection strategies). These results also highlight the need to develop methodologies that can improve our understanding of the mapping between spatial network dependencies and traffic operations.

Traditional traffic models have focused on a detailed description of within-link vehicular interactions. There is limited analytical differentiable work that accounts for between-link interactions.

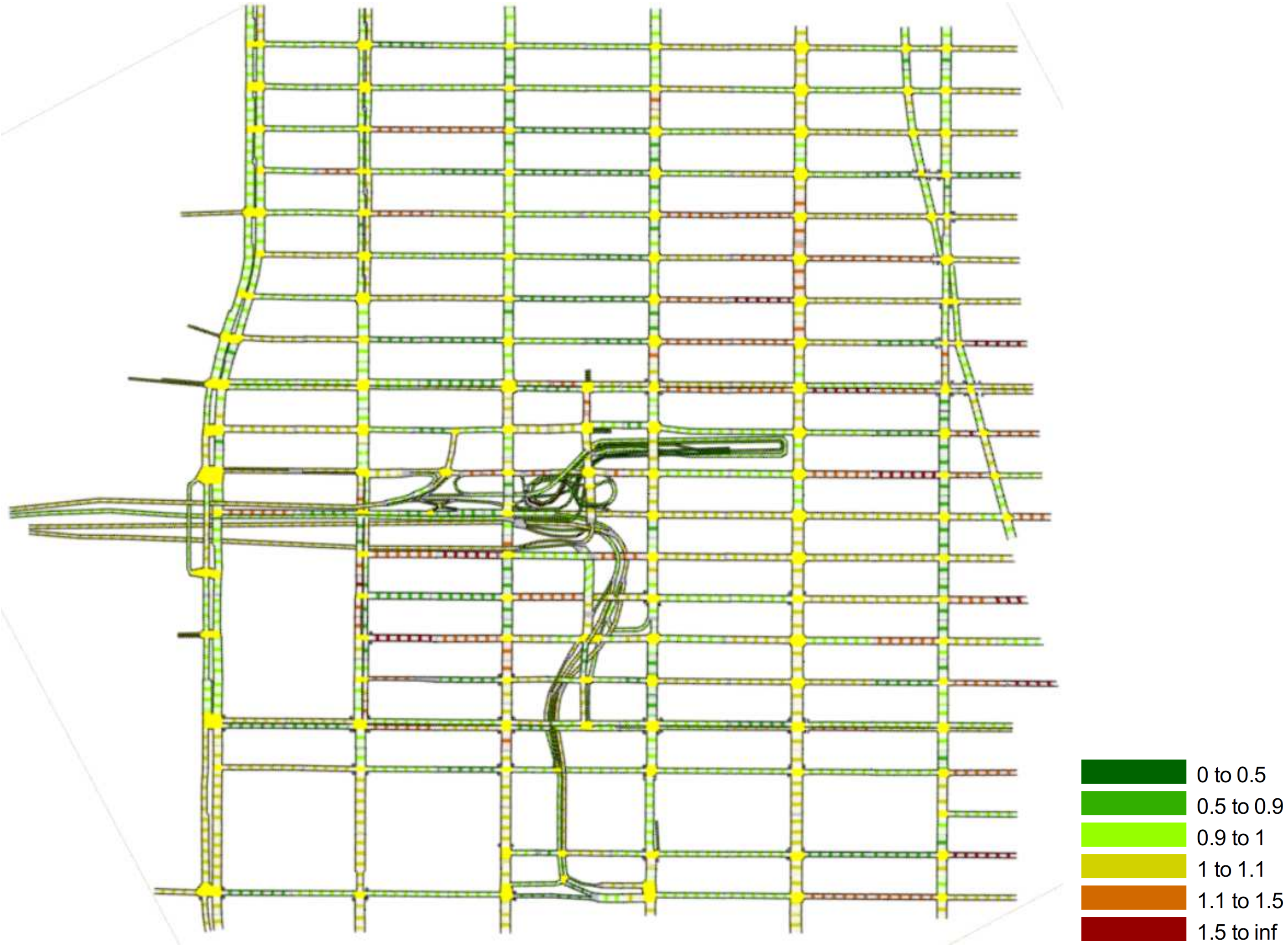


Figure 8: Ratio of the average link density under the proposed plan and that under the existing plan

Recent analytical work that addresses this challenge uses between-queue interaction information to approximate full path or full network distributions of the main performance measures (Osorio and Yamani; 2016; Flötteröd and Osorio; 2014; Osorio and Wang; 2016). In particular, to formulate models that are suitable for large-scale network optimization, the work of Osorio and Yamani (2016); Osorio and Wang (2016) illustrate how link boundary conditions can be described aggregately, i.e., without the need of a detailed description of the within-link state. The complexity of such models is linear, rather than exponential, in the number of links in the network and is independent of the space capacity of the links. This makes them suitable for large-scale network optimization. The use of such detailed models to perform large-scale SO is a topic of ongoing research.

Acknowledgments

The work of C. Osorio was partially funded by National Science Foundation under Grant No.1351512. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The work of X. Chen was partially funded by the Foundation for Science and Technology (FCT) through the MIT Portugal Program (PhD grant SFRH/BD/51296/2010) and by the European Regional Development Fund (FEDER) through the grant EMSURE - Energy and Mobility for Sustainable Regions (CENTRO-07-0224-FEDER-002004). We are grateful to Linsen Chong and Evan Fields for their help in creating Figures 7 and 8.

A SO algorithm

This SO algorithm is formulated in detail in Osorio and Bierlaire (2013) is based on the derivative-free trust region algorithm of Conn et al. (2009). The notation used is that of Osorio and Bierlaire (2013). The parameters of the algorithm are set according to the values in Osorio and Bierlaire (2013).

0. **Initialization.**

   Define for a given iteration k: $m_k(x, y; \alpha_k, \beta_k, q)$ as the metamodel (denoted hereafter as $m_k(x)$), x_k as the iterate, Δ_k as the trust region radius, $\nu_k = (\alpha_k, \beta_k)$ as the vector of parameters of m_k, n_k as the total number of simulation runs carried out up until and including iteration k, u_k as the number of successive trial points rejected, ε_k as the measure of stationarity (norm of the derivative of the Lagrangian function of the trust region (TR) subproblem with regards to the endogenous variables) evaluated at x_k.

The constants $\eta_1, \gamma, \gamma_{inc}, \varepsilon_c, \bar{\tau}, \bar{d}, \bar{u}, \Delta_{max}$ are given such that: $0 < \eta_1 < 1$, $0 < \gamma < 1 < \gamma_{inc}$, $\varepsilon_c > 0$, $0 < \bar{\tau} < 1$, $0 < \bar{d} < \Delta_{max}$, $\bar{u} \in \mathbb{N}^*$. Set the total number of simulation runs permitted (across all points) n_{max}, this determines the computational budget. Set the number of simulation replications per point $\tilde{r}$ (here we use $\tilde{r} = 1$).

Set $k = 0, n_0 = 1, u_0 = 0$. Determine x_0 and Δ_0 ($\Delta_0 \in (0, \Delta_{max}]$).

Given the initial point x_0, compute $f_A(x_0)$ (analytical approximation of Eq. (7)) and $\hat{f}(x_0)$ (simulated estimate of Eq. (7)), fit an initial model m_0 (i.e., compute ν_0).

1. **Criticality step.** If $\varepsilon_k \leq \varepsilon_c$, then switch to *conservative mode.*

2. **Step calculation.** Compute a step s_k that reduces the model m_k and such that $x_k + s_k$ (the trial point) is in the trust region (i.e. approximately solve the TR subproblem).

3. **Acceptance of the trial point.** Compute $\hat{f}(x_k + s_k)$ and

$$\rho_k = \frac{\hat{f}(x_k) - \hat{f}(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

   - If $\rho_k \geq \eta_1$, then accept the trial point: $x_{k+1} = x_k + s_k,\ u_k = 0$.
   - Otherwise, reject the trial point: $x_{k+1} = x_k,\ u_k = u_k + 1$.

   Include the new observation in the set of sampled points ($n_k = n_k + \tilde{r}$), and fit the new model m_{k+1}.

4. **Model improvement.** Compute $\tau_{k+1} = \frac{\|\nu_{k+1} - \nu_k\|}{\|\nu_k\|}$. If $\tau_{k+1} < \bar{\tau}$, then improve the model by simulating the performance of a new point x, which is uniformly drawn from the feasible space. Evaluate f_A and $\hat{f}$ at x. Include this new observation in the set of sampled points ($n_k = n_k + \tilde{r}$). Update m_{k+1}.

5. **Trust region radius update.**

$$\Delta_{k+1} = \begin{cases} \min\{\gamma_{inc}\Delta_k, \Delta_{max}\} & \text{if } \rho_k > \eta_1 \\ \max\{\gamma\Delta_k, \bar{d}\} & \text{if } \rho_k \leq \eta_1 \text{ and } u_k \geq \bar{u} \\ \Delta_k & \text{otherwise.} \end{cases}$$

   If $\rho_k \leq \eta_1$ and $u_k \geq \bar{u}$, then set $u_k = 0$.

   If $\Delta_{k+1} \leq \bar{d}$, then switch to *conservative mode.*

Set $n_{k+1} = n_k, u_{k+1} = u_k$, $k = k + 1$.
If $n_k < n_{max}$, then go to Step 1. Otherwise, stop.

References

Abu-Lebdeh, G. and Benekohal, R. (2003). Design and evaluation of dynamic traffic management strategies for congested conditions, *Transportation Research Part A* **37**(2): 109–127.

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

Ben-Akiva, M., Cuneo, D., Hasan, M., Jha, M. and Yang, Q. (2003). Evaluation of freeway control using a microscopic simulation laboratory, *Transportation Research Part C* **11**(1): 29–50.

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

Branke, J., Goldate, P. and Prothmann, H. (2007). Actuated traffic signal optimization using evolutionary algorithms, *Proceedings of the 6th European Congress and Exhibition on Intelligent Transport Systems and Services*.

Bullock, D., Johnson, B., Wells, R. B., Kyte, M. and Li, Z. (2004). Hardware-in-the-loop simulation, *Transportation Research Part C* **12**(1): 73 – 89.

Chen, X., Osorio, C., Marsico, M., Talas, M., Gao, J. and Zhang, S. (2015). Simulation-based adaptive traffic signal control algorithm, *Transportation Research Board Annual Meeting*, Washington DC, USA.

Chen, X., Osorio, C. and Santos, B. F. (2012). A simulation-based approach to reliable signal control, *Proceedings of the International Symposium on Transportation Network Reliability (INSTR)*.

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

Chow, A. H. F. and Lo, H. K. (2007). Sensitivity analysis of signal control with physical queuing: Delay derivatives and an application, *Transportation Research Part B* **41**(4): 462–477.

Conn, A. R., Scheinberg, K. and Vicente, L. N. (2009). Global convergence of general derivative-free trust-region algorithms to first- and second-order critical points, *SIAM Journal on Optimization* **20**(1): 387–415.

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.

Diakaki, C., Papageorgiou, M. and Aboudolas, K. (2002). A multivariable regulator approach to traffic-responsive network-wide signal control, *Control Engineering Practice* **10**: 183–195.

Flötteröd, G. and Osorio, C. (2014). Stochastic analytical dynamic queueing network model with spill-back, *International Symposium on Dynamic Traffic Assignment (DTA).*

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

Geroliminis, N. and Skabardoni, A. (2011). Identification and analysis of queue spillovers in city street networks, *IEEE Transactions on Intelligent Transportation Systems* **12**(4): 1107 – 1115.

Gregoire, J., Qian, X., Frazzoli, E., de La Fortelle, A. and Wongpiromsarn, T. (2014). Capacity-aware back-pressure traffic signal control, *IEEE Trans. Control of Networked Systems.* forthcoming.

Hale, D. (2005). Traffic network study tool TRANSYT-7F, *Technical report*, McTrans Center in the University of Florida, Gainesville, Florida.

Heidemann, D. (1994). Queue length and delay distributions at traffic signals, *Transportation Research Part B* **28**(5): 377–389.

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

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

Jain, R. and Smith, J. M. (1997). Modeling vehicular traffic flow using M/G/C/C state dependent queueing models, *Transportation Science* **31**(4): 324–336.

Joshi, S., Rathi, A. and Tew, J. (1995). An improved response surface methodology algorithm with an application to traffic signal optimization for urban networks, *in* C. Alexopoulos, K. Kang, W. R. Lilegdon and D. Goldsman (eds), *Proceedings of the 1995 Winter Simulation Conference*, pp. 1104–1109.

Lebacque, J. and Khoshyaran, M. (2005). First–order macroscopic traffic flow models: intersection modeling, network modeling, *in* H. Mahmassani (ed.), *Proceedings of the 16th International Symposium on Transportation and Traffic Theory*, Elsevier, Maryland, USA, pp. 365–386.

Li, P., Abbas, M., Pasupathy, R. and Head, L. (2010). Simulation-based optimization of maximum green setting under retrospective approximation framework, *Transportation Research Record* **2192**: 1–10.

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

Lioris, J., Kurzhanskiy, A. A., Triantafyllos, D. and Varaiya, P. (2014). Control experiments for a network of signalized intersections using the '.Q' simulator, *IFAC/IEEE Workshop on Discrete Event Systems*, Cachan, France.

MapQuest.com, Inc (2015). New York City, NY, Scale undetermined; generated by Xiao Chen; using"MapQuest.com, Inc", `http://www.mapquest.com`. Accessed: 2015-01-02.

Michalopoulos, P. G. and Stephanopoulos, G. (1977a). Oversaturated signal systems with queue length constraints I: Single intersection, *Transportation Research* **11**(6): 413 – 421.

Michalopoulos, P. G. and Stephanopoulos, G. (1977b). Oversaturated signal systems with queue length constraints II: Systems of intersections, *Transportation Research* **11**(6): 423 – 428.

NYCDEP (2006). City Tunnel No. 3, Stage 2 Manhattan Leg Shaft 33B Project, `http://www.nyc.gov/html/dep/pdf/shaft33b/4-10transit.pdf`. Accessed: 2015-01-05.

NYCDOT (2010). Queensboro Bridge Bus Priority Study, `http://www.nyc.gov/html/brt/downloads/pdf/201110_qbb_approach_summary.pdf`. Accessed: 2015-01-05.

NYCDOT (2011). Queensboro bridge bus priority study, `http://www.nyc.gov/html/brt/html/other/queensboro.shtml`. Accessed: 2015-08-14.

NYCDOT (2012). 2012 New York City Bridge Traffic Volumes, `http://www.nyc.gov/html/dot/downloads/pdf/bridge-traffic-report-2012.pdf`. Accessed: 2015-03-1.

NYCDOT (2015). NYCDOT Bridges, `http://www.nyc.gov/html/dot/html/infrastructure/bridges.shtml`. Accessed: 2015-01-05.

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

Osorio, C. and Bierlaire, M. (2009). An analytic finite capacity queueing network model capturing the propagation of congestion and blocking, *European Journal of Operational Research* **196**(3): 996–1007.

Osorio, C. and Bierlaire, M. (2013). A simulation-based optimization framework for urban transportation problems, *Operations Research* **61**(6): 1333–1345.

Osorio, C., Chen, X., Gao, J., Talas, M. and Marsico, M. (2015). On the control of highly congested urban networks with intricate traffic patterns: a New York City case study, *Technical report*, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT). Available at: http://web.mit.edu/osorioc/www/papers/osoChenNYCDOTOfflineSO.pdf .

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

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

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

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

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

Osorio, C. and Selvam, K. (2016). Simulation-based optimization: achieving computational efficiency through the use of multiple simulators, *Transportation Science* . Forthcoming. Available at: http://web.mit.edu/osorioc/www/papers/osoSelMultiModel.pdf.

Osorio, C. and Wang, C. (2016). On the analytical approximation of joint aggregate queue-length distributions for traffic networks: a stationary finite capacity Markovian network approach, *Transportation Research Part B* . Forthcoming. Available at: http://web.mit.edu/osorioc/www/papers/osoWangAggDisagg.pdf .

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

Osorio, C., Zhang, C. and Flötteröd, G. (2016). Efficient calibration techniques for large-scale traffic simulators, *submitted for journal publication* . Available at: http://web.mit.edu/osorioc/www/papers/osoZhaFlo16Calib.pdf .

Papageorgiou, M., Diakaki, C., Dinopoulou, V., Kotsialos, A. and Wang, Y. (2003). Review of road traffic control strategies, *Proceedings of the IEEE* **91**(12): 2043–2067.

Papageorgiou, M. and Varaiya, P. (2009). Link vehicle-count - the missing measurement for traffic control, *in* A. Chassiakos (ed.), *Proceedings of the 12th IFAC Symposium on Control in Transportation Systems*, Redondo Beach CA, USA.

Park, B., Yun, I. and Ahn, K. (2009). Stochastic optimization for sustainable traffic signal control, *International Journal of Sustainable Transportation* **3**(4): 263–284.

Rathi, A. K. (1988). A control scheme for high traffic density sectors, *Transportation Research Part B* **22B**(2): 81–101.

Skabardonis, A. . and Geroliminis, N. (2008). Real-time monitoring and control on signalized arterials, *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* **12**(2): 6474.

Spall, J. C. and Chin, D. C. (1997). Traffic-responsive signal timing for system-wide traffic control, *Transportation Research Part C* **5**(3-4): 153–163.

Stafford, R. (2006). *The Theory Behind the 'randfixedsum' Function.* http://www.mathworks.com/matlabcentral/fileexchange/9700.

Stevanovic, A., Stevanovic, J., Zhang, K. and Batterman, S. (2009). Optimizing traffic control to reduce fuel consumption and vehicular emissions, *Transportation Research Record* **2128**: 105–113.

Stevanovic, J., Stevanovic, A., Martin, P. T. and Bauer, T. (2008). Stochastic optimization of traffic control and transit priority settings in VISSIM, *Transportation Research Part C* **16**(3): 332 – 349.

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

Tanner, J. C. (1962). A theoretical analysis of delays at an uncontrolled intersection, *Biometrika* **49**: 163–170.

TSS (2013). *AIMSUN 7 Dynamic Simulators User's Manual*, Transport Simulation Systems.

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

Wongpiromsarn, T., Uthaicharoenpong, T., Wang, Y., Frazzoli, E. and Wang, D. W. (2012). Distributed traffic signal control for maximum network throughput, *IEEE Conference on Intelligent Transportation Systems*, p. 588595.

Xin, W., Chang, J., Muthuswamy, S. and Talas, M. (2013). "Midtown in Motion": a new active traffic management methodology and its implementation in New York City, *Transportation Research Board Annual Meeting*, Washington DC, USA.

Yun, I. and Park, B. (2006). Application of stochastic optimization method for an urban corridor, *Proceedings of the Winter Simulation Conference*, pp. 1493–1499.