CARA: A CONGESTION-AWARE ROUTING ALGORITHM FOR AUTONOMOUS MOBILITY-ON-DEMAND SYSTEMS

Mauro Salazar

Matthew Tsao

Izabel Aguiar

Maximilian Schiffer, Ph.D.

Marco Pavone, Ph.D.

Current Word Count: 4190; Intro: 1041; Model: 1663; Results: 1118; Conclusion: 305

Submission Date: August 15, 2018

Abstract

We present a congestion-aware routing policy for Autonomous Mobility-on-Demand (AMoD) systems that accounts for the impact of road traffic on travel time. Specifically, we develop a congestion-aware routing algorithm (CARA) that captures road utilization dependent travel times via a piecewise affine approximation of the Bureau of Public Roads (BPR) model. Such an approximation largely retains the key features of the BPR model, while enabling the design of a real-time, convex quadratic optimization algorithm to compute congestion-aware routes for an AMoD fleet. Through a real-world case study of Manhattan, we compare CARA to existing routing approaches, the first one congestion-unaware and the second one based on a threshold congestion model. Numerical results show that CARA significantly outperforms the state of the art, with improvements in terms of travel time and global cost in the order of 20%, and computation times compatible with a real-time implementation.

Keywords: Network Flow Models, Autonomous Mobility, Congestion

Current Word Count: 4190; Intro: 1041; Model: 1663; Results: 1118; Conclusion: 305

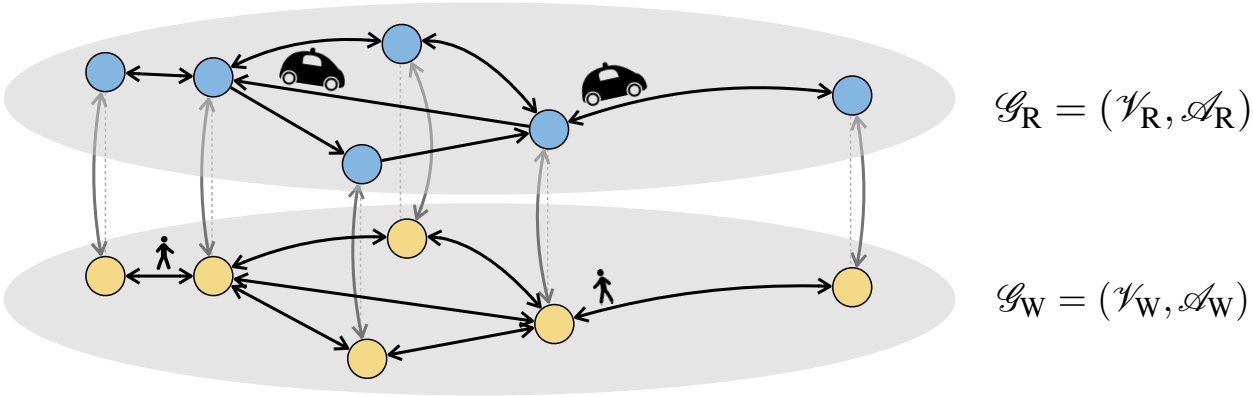

FIGURE 1 The AMoD network. The colored dots represent intersections and the black arrows denote road links or pedestrian pathways. The gray dotted lines highlight geographically equivalent nodes connected by gray mode-switching arcs.

1. INTRODUCTION

Congestion remains a central problem in today's transportation systems, especially in densely populated urban areas. While congestion phenomena have been attenuated by subsidiary modes of transportation (e.g., public transport) in the past, in recent years congestion-related problems have rapidly increased. Even mature cities are struggling with these problems, with the sustainability of mobility systems restricted by current infrastructure and space limitations. Additionally, current societal trends increase congestion and emphasize its negative impacts. First, steadily increasing urbanization leads to higher population densities and thus higher mobility demand in cities (*1*). Second, commuters' individual mobility needs and comfort preferences have led to the decrease in the utilization of public transportation (*2*). Third, mobility-on-demand services such as Uber and Lyft are rapidly growing as an alternative to public transportation and individual car ownership (*3*). Consequently, (selfish) traffic on roads is steadily growing and increases congestion even further. Current transport Key Performance Indicators (KPIs) reflect this situation. According to (*4*) the annual delay per commuter exceeds 42 hours and one third of commuter trips are reported as having "extreme, severe, or heavy" congestion. In New York City, the average speed in Manhattan decreased from 6.5 mph to 4.7 mph between 2012 and 2017 (*5*).

Cities face spatial limitations in addressing the issue of congestion as the available infrastructure (e.g., roads, traffic signals, highways, train lines) and its capacities are largely fixed. It is thus necessary to develop more efficient systems in road transport to supplement existing public transportation. Herein, Autonomous Mobility-on-Demand (AMoD) systems represent a promising solution. An AMoD system consists of a fleet of self-driving vehicles designed to carry passengers from their origins to their destinations. As passengers request trips, the central operator of the AMoD system assigns each passenger to an empty vehicle. Once the passengers are dropped off, the central operator computes an optimal rebalancing route for assigning the vehicle to the next request. The central operator is thus in control of simultaneously optimizing the routes for all vehicles in the system. As such, AMoD can replace current forms of mobility-on-demand (e.g., taxis and ride-hailing) while reducing the cost of travel (*6*). Furthermore, because the entire fleet of vehicles in an AMoD system is centrally controlled, this form of autonomous mobility holds promise to reduce congestion.

The potential of AMoD systems to alleviate congestion in cities depends on congestion-aware (CA) routing. CA routing considers the natural capacities of roads and the effects of vehicle flows on travel times through *volume-delay functions*. CA routing is not a new idea: navigation providers such as Google Maps and Waze have incorporated features that allow users to view how real-time and estimated traffic congestion will impact the routes of their commutes (*7, 8*). CA routing with conventional algorithms, however, can only *passively suggest* the routing of a single vehicle to avoid traffic, and furthermore only *anticipate* the traffic from other selfish vehicles. Conversely, using CA routing in an AMoD setting allows one to *actively control* the routes of all vehicles in the fleet under complete system information. With the aim to enable realistic CA routing for AMoD systems, we develop and study the use of a volume-delay function designed for convex optimization purposes. To set this work apart from the status quo, we briefly review related literature and state our aims and scope of the paper.

1.1 Related Literature

The most widely-used volume-delay function is the one developed by the Bureau of Public Roads (BPR) (*9*), although many other "BPR-type" functions exist (*10*). In related work these functions have been used for algorithmic approaches to dynamic estimation of congestion on a network (*11*) and CA route planning in agent-based models (*12, 13*). Furthermore, CA route planning of AMoD systems has been simulated in (*14–18*), and optimized for dynamic traffic assignment in (*19, 20*). These approaches have provided CA routing *analysis*, but are limited to simulations and lack control algorithms for both passenger requests and vehicle rebalancing. Thus far, the CA *control* of AMoD systems has been limited to thresholded approximations of the BPR function (*21–23*). In this approximation the time required to travel on a road is defined through a thresholded approach: the cars on a road are permitted to travel at a free-flow speed if the *capacity* of the road has not yet been reached. Additional cars beyond this capacity, however, make this road impossible to traverse. Although this model provides a conservative approach to capturing the effect of congestion on travel time, it oversimplifies congestion phenomena and may lead to suboptimal route patterns. To the best of our knowledge, no algorithmic framework for CA routing of AMoD systems currently exists that allows one to simultaneously address a) the more precise BPR-defined effects of congestion and b) real-time optimization of routing decisions for customer requests and rebalancing vehicles.

1.2 Aims and Scope

To resolve the drawbacks outlined above, we propose a congestion-aware routing algorithm (CARA) that leverages a piecewise-affine approximation of the BPR congestion model (*9*). We study the impact of such an algorithm in a real-world case study of Manhattan. We compare CARA to two different baselines, the first one congestion unaware and the second one capturing congestion via a threshold model. We study CA traffic routing in Manhattan and show that CARA significantly improves the state of the art with respect to travel times and global cost, while featuring computation times compatible with a real-time implementation. It should be noted that, while the BPR function does not capture every effect of congestion (e.g., spillback, heterogeneous vehicles, or intersection delays), it is a well-accepted *model* which suits the aim of this paper, namely, not to perfectly capture traffic dynamics, but to approximate them precisely enough for optimization and control purposes.

1.3 Organization

The remainder of this paper is structured as follows: in Section 2 we detail the methodology and develop CARA: a flow-based optimization framework for CA routing of AMoD systems. In Section 3 we numerically evaluate the performance of CARA through a case study of AMoD traffic routing in Manhattan. We conclude the paper in Section 4 with a discussion of our results and an outlook on future research directions.

2. METHODOLOGY

This section provides the methodological foundation for CARA. We introduce a multi-commodity flow model to represent the physical constraints of the transportation system and AMoD fleet in Section 2.1. In Section 2.2 we specify an objective function for the model with the goal to optimize social welfare. We present an approximation of the BPR volume-delay function to consider congestion while preserving model convexity in Section 2.3, and conclude in Section 2.4 with a brief discussion.

2.1 Multi-commodity Flow Based Optimization Approach

Recall from Section 1.2, that we aim to model a network with two modes of transportation: walking and riding AMoD. This transportation network can be modeled on a digraph $\mathscr{G} = (\mathscr{V}, \mathscr{A})$ as shown in Fig. 1. The graph consists of a set of vertices $\mathscr{V}$ and a set of arcs $\mathscr{A} \subseteq \mathscr{V} \times \mathscr{V}$. To capture both modes of transportation, $\mathscr{G}$ comprises a road network layer $\mathscr{G}_\mathrm{R} = (\mathscr{V}_\mathrm{R}, \mathscr{A}_\mathrm{R})$ and a walking layer $\mathscr{G}_\mathrm{W} = (\mathscr{V}_\mathrm{W}, \mathscr{A}_\mathrm{W})$. The road network layer represents intersections $i \in \mathscr{V}_\mathrm{R}$ and road links $(i,j) \in \mathscr{A}_\mathrm{R}$, while the pedestrian layer models walkable streets $(i,j) \in \mathscr{A}_\mathrm{W}$ in between intersections $i \in \mathscr{V}_\mathrm{W}$. Whereas the locations of nodes and arcs in $\mathscr{G}_\mathrm{R}$ and $\mathscr{G}_\mathrm{W}$ may coincide geographically, we maintain a distinction between walking and riding in an AMoD vehicle. Additionally, switching arcs out of set $\mathscr{A}_\mathrm{C} \subseteq \mathscr{V}_\mathrm{R} \times \mathscr{V}_\mathrm{W}$ connect the pedestrian layer to the road network layer, and model the customer's ability to switch transportation modes by hailing an AMoD ride or exiting a car. Collectively, it holds $\mathscr{V} = \mathscr{V}_\mathrm{W} \cup \mathscr{V}_\mathrm{R}$ and $\mathscr{A} = \mathscr{A}_\mathrm{W} \cup \mathscr{A}_\mathrm{R} \cup \mathscr{A}_\mathrm{C}$.

Each arc has a specific length d_{ij} and a constant nominal travel time t_{ij}^N denoting the walking time for arcs $(i,j) \in \mathscr{A}_\mathrm{W}$, the time to hail or exit an AMoD vehicle for arcs $(i,j) \in \mathscr{A}_\mathrm{C}$, and the travel time under free-flow conditions (without traffic) for arcs $(i,j) \in \mathscr{A}_\mathrm{R}$. As in (*22*), we model the energy consumption of AMoD vehicles, assuming a constant nominal speed $v_{ij} = \frac{d_{ij}}{t_{ij}^N}$ for each arc. Furthermore, we assume the AMoD fleet to be composed of lightweight electric vehicles with an overall efficiency η_EV and full recuperation capabilities. Thus the energy consumption per road arc is

$$e_{ij} = \left(\frac{\rho_\mathrm{a}}{2} \cdot A_\mathrm{f} \cdot c_\mathrm{d} \cdot v_{ij}^2 + c_\mathrm{r} \cdot m_\mathrm{v} \cdot g \right) \cdot \frac{d_{ij}}{\eta_\mathrm{EV}} \; (i,j) \in \mathscr{A}_\mathrm{R}, \tag{1}$$

where the aerodynamic drag is determined by the air density ρ_a, the frontal area A_f, and the drag coefficient c_d, and the friction of the wheels on the road is determined by the rolling friction coefficient c_r, the mass of the vehicle m_v, and the gravitational acceleration g (*24*).

Each travel demand $m \in \mathscr{M} = \{1, \ldots, M\}$ consists of an origin destination pair (o_m, d_m) on the walking digraph $\mathscr{G}_\mathrm{W}$ and a demand rate α_m that denotes the number of customers that are requesting the same trip per unit time. To trace customer flows, $f_m(i,j)$ denotes the flow of customers on arc $(i,j) \in \mathscr{A}$ for demand $m \in \mathscr{M}$. As AMoD vehicles may need to relocate between two customer requests, $f_0(i,j)$ denotes the rebalancing flow of empty vehicles on $(i,j) \in \mathscr{A}_\mathrm{R}$.

This notation is sufficient to derive a basic multi-commodity flow model for our planning

problem at hand. Consider the cost function J mapping the set of flows $\{f_m(\cdot,\cdot)\}_m, f_0(\cdot,\cdot)$ into the set of non-negative real numbers $\mathbb{R}_{\geq 0}$. We state the AMoD optimal routing problem as

$$\min_{\{f_m(\cdot,\cdot)\}_m, f_0(\cdot,\cdot)} \quad J\big(\{f_m(\cdot,\cdot)\}_m, f_0(\cdot,\cdot)\big) \tag{2a}$$

s.t.

$$\sum_{i:(i,j)\in\mathscr{A}} f_m(i,j) + \mathbb{1}_{j=o_m}\cdot\alpha_m = \sum_{k:(j,k)\in\mathscr{A}} f_m(j,k) + \mathbb{1}_{j=d_m}\cdot\alpha_m \qquad \forall m\in\mathscr{M},\, j\in\mathscr{V} \tag{2b}$$

$$\sum_{i:(i,j)\in\mathscr{A}_{\mathrm{R}}} \left(f_0(i,j) + \sum_{m\in\mathscr{M}} f_m(i,j)\right) = \sum_{k:(j,k)\in\mathscr{A}_{\mathrm{R}}} \left(f_0(j,k) + \sum_{m\in\mathscr{M}} f_m(j,k)\right) \qquad \forall j\in\mathscr{V}_{\mathrm{R}} \tag{2c}$$

$$f_m(i,j) \geq 0 \qquad \forall (i,j)\in\mathscr{A} \tag{2d}$$

$$f_0(i,j) \geq 0 \qquad \forall (i,j)\in\mathscr{A}_{\mathrm{R}}, \tag{2e}$$

where $\mathbb{1}_x$ is a boolean indicator function. Linear constraints (2b) and (2c) ensure that the mass of customers and vehicles, respectively, are conserved on every road node. Inequality constraints (2d) and (2e) ensure that the customer and rebalancing flows are non-negative.

2.2 AMoD Objective

We use the cost function (2a) to model the social cost of serving the transportation requests in a similar fashion as (*22*). Specifically, we aim to minimize the total travel time and the operational costs of the AMoD system. We assume customers to have the same value of time V_{T}. We separate the cost of operating the AMoD fleet into a distance-dependent cost V_{D} due to depreciation and maintenance, and an energy-consumption-dependent cost V_{E}. Collectively, the cost function (2a) is

$$\begin{aligned} J_{\mathrm{M}}\big(\{f_m(\cdot,\cdot)\}_m, f_0(\cdot,\cdot)\big) =& V_{\mathrm{T}}\cdot \sum_{m\in\mathscr{M},(i,j)\in\mathscr{A}} t_{ij}\big(f(i,j)\big)\cdot f_m(i,j) \\ &+ \sum_{(i,j)\in\mathscr{A}_{\mathrm{R}}} (V_{\mathrm{D}}\cdot d_{ij} + V_{\mathrm{E}}\cdot e_{ij})\cdot f(i,j), \end{aligned} \tag{3}$$

where $f(i,j) = f_0(i,j) + \sum_{m\in\mathscr{M}} f_m(i,j)$ is the total flow on arc (i,j). The travel time t_{ij} on road arcs is modeled as a function of road usage, while $t_{ij} = t_{ij}^N$ gives the time to walk or switch transportation layer. For road arcs we use the BPR volume-delay function (*9*)

$$t_{ij}\big(f(i,j)\big) = t_{ij}^N \cdot F_{\mathrm{BPR}}\left(\frac{f(i,j)+u_{ij}^{\mathrm{R}}}{c_{ij}^{\mathrm{R}}}\right) \qquad \forall (i,j)\in\mathscr{A}_{\mathrm{R}}, \tag{4}$$

where u_{ij}^{R} is the exogenous road usage caused by, e.g., the presence of private cars, c_{ij}^{R} being the nominal road capacity, and

$$F_{\mathrm{BPR}}(x) = 1 + 0.15\cdot x^4. \tag{5}$$

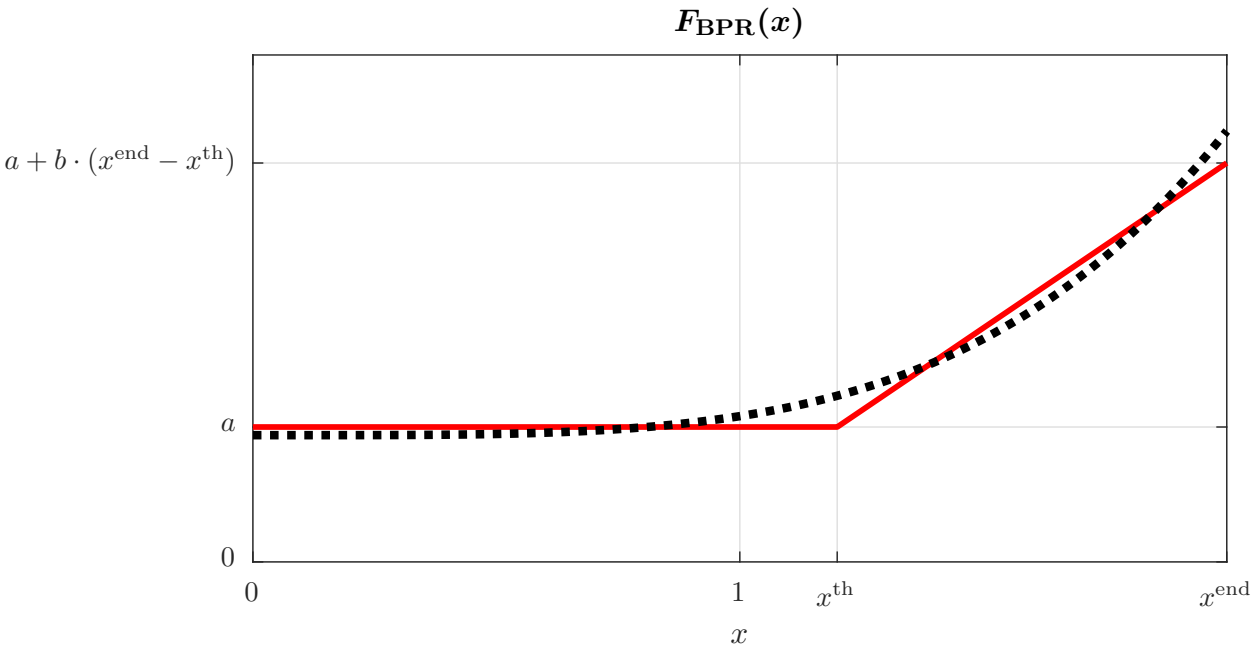

FIGURE 2 The BPR function $F_{\mathrm{BPR}}(x)$ (black dotted) and its piecewise affine fit (red solid).

2.3 BPR Model Approximation

We aim to minimize (2a) subject to the constraints specified by (2b)-(2e). However, J_{M} is a non-convex polynomial of the decision variables, namely the vehicle flows $\{f_m(\cdot,\cdot)\}_m, f_0(\cdot,\cdot)$. Hence, this model is, in general, computationally intractable, as there does not exist a known algorithm to reliably and efficiently solve large-scale non-convex polynomial optimization problems. To resolve this non-convexity we discuss a piecewise affine approximation to the BPR function.

The non-convexity arises from the product $t_{ij}(f(i,j)) \cdot f_m(i,j)$. Hence, we approximate this term with a convex function to ensure scalability to large-size problem instances. Specifically, we fit Eq. (5) using a piecewise affine approximation as shown in Fig. 2:

$$y = \begin{cases} a & \text{if } x \in [0, x^{\mathrm{th}}] \\ a + b \cdot (x - x^{\mathrm{th}}) & \text{if } x \in (x^{\mathrm{th}}, x^{\max}] \end{cases}, \tag{6}$$

where a is the height of the horizontal line, b the slope of the second line, x^{th} is the non-smooth threshold in the piecewise affine approximation, and $x^{\max}$ defines the approximation window. This way we approximate the BPR function (4) as

$$t_{ij} = \begin{cases} t_{ij}^N \cdot a & \text{if } f(i,j) \le c_{ij}^{\mathrm{R,th}} \\ t_{ij}^N \cdot \left(a + b \cdot \frac{f(i,j) - c_{ij}^{\mathrm{R,th}}}{c_{ij}^{\mathrm{R}}} \right) & \text{if } f(i,j) > c_{ij}^{\mathrm{R,th}} \end{cases} =: t_{ij}^0 + \tau_{ij} \cdot \varepsilon(i,j) \ , \tag{7}$$

where $t_{ij}^0 = a \cdot t_{ij}^N$, $\tau_{ij} = b \cdot t_{ij}^N / c_{ij}^{\mathrm{R}}$, and the slack variable $\varepsilon(i,j)$ denotes how much the total flow on arc $(i,j) \in \mathscr{A}_{\mathrm{R}}$ exceeds its threshold capacity $c_{ij}^{\mathrm{R,th}} = x^{\mathrm{th}} \cdot c_{ij}^{\mathrm{R}}$, that is,

$$\varepsilon(i,j) = \max\left\{0, f(i,j) + u_{ij}^{\mathrm{R}} - c_{ij}^{\mathrm{R,th}}\right\}. \tag{8}$$

Provided that J is a decreasing function of $\varepsilon(i,j)$, this piecewise affine cost can be represented by

the following linear inequality constraints

$$\varepsilon(i,j) \geq f(i,j) + u_{ij}^{\mathrm{R}} - c_{ij}^{\mathrm{R,th}} \tag{9}$$

$$\varepsilon(i,j) \geq 0. \tag{10}$$

Hence, the total travel time ζ on road links can be expressed as

$$\begin{aligned}
\zeta := T^{\mathrm{R}} \cdot \sum_{m\in\mathscr{M}} \alpha_m &= \sum_{(i,j)\in\mathscr{A}_{\mathrm{R}}} t_{ij}\big(f(i,j)\big) \cdot \sum_{m\in\mathscr{M}} f_m(i,j) \\
&= \sum_{(i,j)\in\mathscr{A}_{\mathrm{R}}} t_{ij}^0 \cdot \sum_{m\in\mathscr{M}} f_m(i,j) + \sum_{(i,j)\in\mathscr{A}_{\mathrm{R}}} \tau_{ij} \cdot \varepsilon(i,j) \cdot \sum_{m\in\mathscr{M}} f_m(i,j) \\
&= \sum_{(i,j)\in\mathscr{A}_{\mathrm{R}}} t_{ij}^0 \cdot \sum_{m\in\mathscr{M}} f_m(i,j) + \sum_{(i,j)\in\mathscr{A}_{\mathrm{R}}} \tau_{ij} \cdot \varepsilon(i,j) \cdot \Big(\varepsilon(i,j) + c_{ij}^{\mathrm{R,th}} - u_{ij}^{\mathrm{R}} - f_0(i,j)\Big) \\
&\leq \sum_{(i,j)\in\mathscr{A}_{\mathrm{R}}} t_{ij}^0 \cdot \sum_{m\in\mathscr{M}} f_m(i,j) + \sum_{(i,j)\in\mathscr{A}_{\mathrm{R}}} \tau_{ij} \cdot \Big(\varepsilon(i,j)^2 + \varepsilon(i,j) \cdot (c_{ij}^{\mathrm{R,th}} - u_{ij}^{\mathrm{R}})\Big) \\
&=: \widehat{\zeta}
\end{aligned} \tag{11}$$

We would like to include the total trip time ζ in the objective for the optimization, but due to the bilinear terms $-\varepsilon(i,j) f_0(i,j)$, ζ is not a convex function of the flow variables. As an alternative, we find a convex approximation of ζ to include in the objective function. Since $\varepsilon(i,j) \cdot f_0(i,j) \geq 0$, removing these bilinear terms makes the expression larger, hence we have $\zeta \leq \widehat{\zeta}$. Without the bilinear terms, $\widehat{\zeta}$ is a convex function of the flow variables, so we include it in the objective function to penalize strategies with long trip times. Considering that the number of rebalancing vehicles has a minor impact with respect to road congestion and converges to zero for perfectly symmetric demand distributions (*21*), $\widehat{\zeta}$ can be used as a metric for the total travel time on road arcs. Specifically, our empirical studies observed that $\frac{|\zeta - \widehat{\zeta}|}{\zeta} \leq 0.1$. In doing so, we approximate the total cost function (2a) with the quadratic bound (11) as

$$\begin{aligned}
J_{\mathrm{Q}}\big(\{f_m(\cdot,\cdot)\}_m, f_0(\cdot,\cdot)\big) = V_{\mathrm{T}} \cdot \underbrace{\left(\sum_{\substack{m\in\mathscr{M},\\ (i,j)\in\mathscr{A}}} t_{ij}^N \cdot f_m(i,j) + \sum_{(i,j)\in\mathscr{A}_{\mathrm{R}}} \tau_{ij} \cdot \Big(\varepsilon(i,j)^2 + \varepsilon(i,j) \cdot (c_{ij}^{\mathrm{R,th}} - u_{ij}^{\mathrm{R}})\Big) \right)}_{\widehat{\zeta}} \\
+ \sum_{(i,j)\in\mathscr{A}_{\mathrm{R}}} (V_{\mathrm{D}} \cdot d_{ij} + V_{\mathrm{E}} \cdot e_{ij}) \cdot f(i,j).
\end{aligned} \tag{12}$$

The AMoD optimal routing problem given by (2) with $J(\cdot,\cdot) = J_{\mathrm{Q}}(\cdot,\cdot)$ subject to (9) then remains a convex quadratic program.

2.4 Discussion

A few comments are in order. First, we assume travel requests to be time-invariant. This assumption is reflected in densely populated urban environments where requests often change slowly compared to the average time needed to complete an individual trip (*25*). Second, we use the BPR function (*9*) to describe the impact of road usage on travel time. While such a function does not perfectly capture traffic phenomena, it is a well-established *model* serving the purpose of designing CA routing algorithms. In order to embed the BPR function in a convex optimization

framework, we approximate it in a piecewise affine fashion. This approximation gives a generalization of simple threshold models that are used in classical traffic flow theory (*26*) and can allow for better mobility service in congested situations, as discussed in the remainder of the paper. Although a piecewise affine function cannot closely approximate a quartic polynomial such as the BPR function on its entire domain, this approximation only needs to be accurate for realistic values of vehicle flows. Furthermore, we relax the piecewise approximation and implement it in the convex optimization framework through a quadratic upper bound. Nevertheless, since the ratio of empty vehicles to passenger carrying vehicles is usually low, this bound is tight enough, as shown in Section 3. Third, CARA captures customer and vehicle routes as fractional flows and does not addresses the stochasticity of the exogenous traffic and customer requests. Arguably, such approximations are acceptable, given the mesoscopic perspective of our study. On the topic of operational algorithms, CARA can be directly extended to operate in real-time. Due to the computational efficiency of the algorithm, CARA can be run periodically with updated real-time information about customer demand to operate in a time varying environment. Furthermore, randomized rounding routing algorithms can compute near-optimal integer-valued flows for individual customers starting from the fractional solution computed by CARA (*27*). The operational implementation of CARA is further discussed in Section 4. Fourth, we assume vehicles to carry only a single customer at a time. This mode of operation is in line with current trends in mobility-on-demand systems such as taxis, Lyft, and Uber. The extension to ride-sharing is a challenging direction for future research. Finally, for the sake of simplicity, we assume all customers to value time and travel comfort in the same way. However, CARA can be extended to capture multiple classes of customers using network flows which are distinct not only in the origin-destination transportation request, but also in the customers' preference profile.

3. COMPUTATIONAL STUDIES

The goal of this section is to assess CARA by evaluating and comparing its performance in a case study to state of the art congestion aware routing algorithms. We begin in Section 3.1 by describing the Manhattan-based case study where we conduct computational studies. In Section 3.2, we compare the performance of CARA to that of a) a congestion-unaware approach and b) an algorithm that captures congestion via a threshold model.

3.1 Manhattan Case Study

Our case study is based on the area of Manhattan in New York City. For this area, we use real data from taxi rides that occurred between 6:00PM and 8:00PM on March 1, 2012 (courtesy of the New York Taxi and Limousine Commission). In total this data set comprises 53,932 taxi rides. Although this number of trips is quite large, it reflects only a fraction of the travel demand between 6:00PM and 8:00PM. In 2017 ride-sharing vehicles used during this time period outnumbered taxis by a ratio of 5:1 (*28*). To reflect this, we dilate the number of requests by a factor of six to emulate the total demand for ride-hailing services in Manhattan during this time window.

We derive the road network from Open Street Map data (*29*) and set the nominal road capacities c^{R}_{ij} proportional to the number of lanes on a road, times its speed limit (*21*). The walking network shows similar spatial characteristics to the road network, but is complementary as all arcs are bi-directional to allow for walking in both directions, even in one-way streets. To account for exogenous privately owned vehicles on the road, we run our simulations for different values of road usage u^{R}, denoting the fraction of a road's free-flow capacities used by private vehicles. For

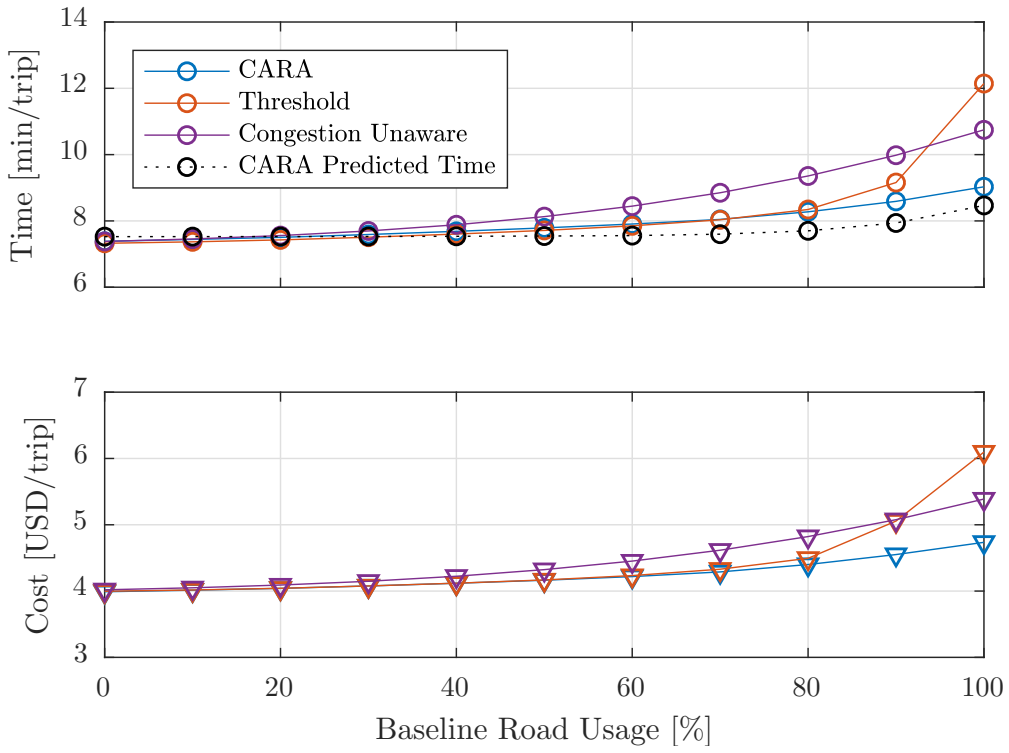

FIGURE 3 Average travel time and cost per passenger trip for all approaches.

a more detailed description of this case study and its parameters we refer to (*22*).

3.2 Results

In the following section, we discuss the results of our case study. We compare CARA to two baselines: a control algorithm using a simple threshold model in which $f(i,j) + u_{ij}^{\mathrm{R}} \leq c_{ij}^{\mathrm{R,th}}$ is enforced (cf. (*21, 22*)), and a naïve congestion-unaware approach in which road utilization does not affect travel times. For the first two models we set $c_{ij}^{\mathrm{R,th}} = x^{\mathrm{th}} \cdot c_{ij}^{\mathrm{R}}$, with $x^{\mathrm{th}} = 1.2$, and we approximate the BPR function with $x^{\mathrm{end}} = 2$ (cf. Fig. 2). We also compute the CARA Predicted Time, which is the average trip time predicted using the piecewise affine function and approximations discussed in section 2.3. For each of the scenarios studied, the computation of the optimal solution took less than four minutes on commodity hardware (Intel Core i7, 16 GB RAM) using Gurobi 7.5.

We measure the quality of CARA with two performance indicators: the average resulting trip time for each solution computed by the BPR function, and the global cost as defined in (3). Fig. 3 and 4 summarize the results. Focusing on these results, we identify different relationships between the performance of all algorithms depending on the exogenous road usage. Additionally, the travel times predicted by CARA are close to those computed by the BPR model, as shown in Fig. 3 and 4.

- $u^{\mathrm{R}} \in [0\%, 20\%] \cdot c^{\mathrm{R}}$: For low levels of road usage (i.e., u^{R} close to zero), all control algorithms show the same performance: in this scenario roads have enough capacity for AMoD vehicles in the absence of private vehicles.
- $u^{\mathrm{R}} \in [40\%, 80\%] \cdot c^{\mathrm{R}}$: As the traffic level increases, the naïve control algorithm performs

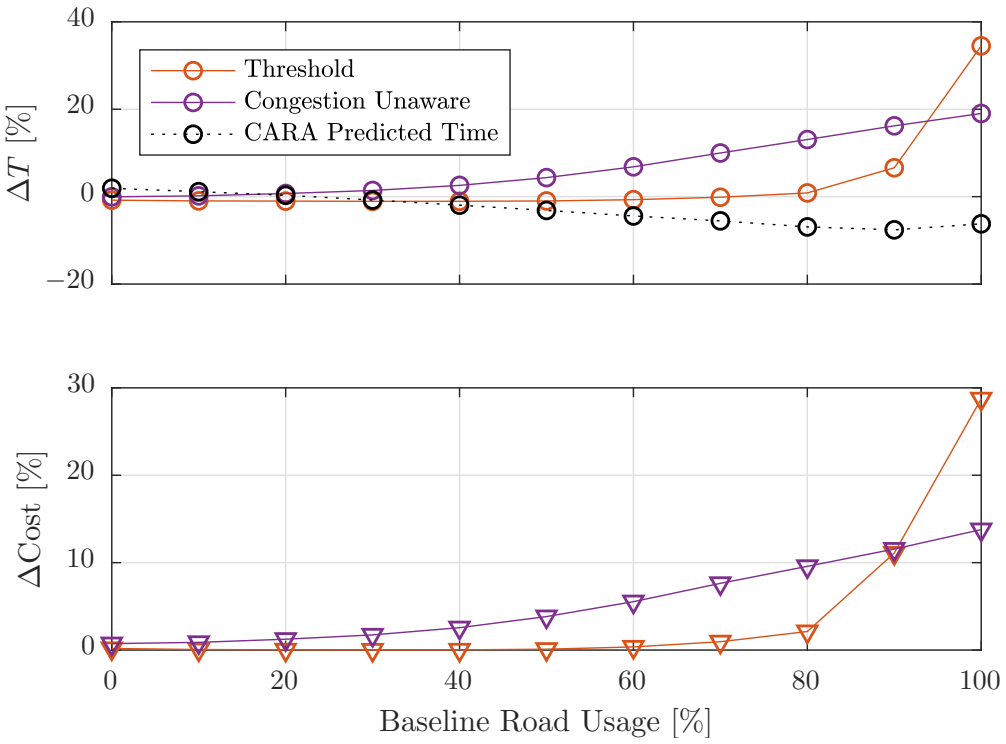

FIGURE 4 Relative deviation of the two baseline solutions compared to CARA. The upper plot also shows the relative error of the CARA predicted time.

worse than the thresholded approach and CARA in terms of both average travel time and operational cost. This shows that the naïve control algorithm, lacking information about a street's capacity, sends the cars on the spatially shortest paths. As a consequence these paths become congested and remain the spatially but not temporally shortest paths due to increased travel times. Conversely, the threshold model and CARA perform similarly better for this range.

- $u^{\mathrm{R}} \in [90\%, 100\%] \cdot c^{\mathrm{R}}$: For heavy levels of traffic CARA significantly outperforms the other two approaches. Especially at 100% exogenous road usage, CARA offers a 20% shorter average trip time than the naïve algorithm and a 35% shorter average trip time than the threshold algorithm. Note that for heavy exogenous road usage, the threshold model performs even worse than the naïve model. This is because the threshold algorithm is not allowed to use roads once the flow on it reaches $c^{\mathrm{R,th}}$. In this case, the only other option is to send customer carrying vehicles on long detours. In reality, $c^{\mathrm{R,th}}$ is not a hard constraint and sending the car along a congested road can still be faster than the aforementioned long detours. Fig. 5 shows the congestion maps for this scenario and the three routing algorithms, where the shortcomings of the congestion unaware and the threshold algorithm are highlighted together with the balanced behaviour of CARA.

In summary, we find that the performance of the threshold congestion model depends heavily on the level of exogenous traffic. In situations of low to medium exogenous traffic, an algorithm using the threshold model can perform much better than a congestion-unaware model. However, under heavy levels of traffic, the threshold model behaves too conservatively, i.e., it underutilizes routes

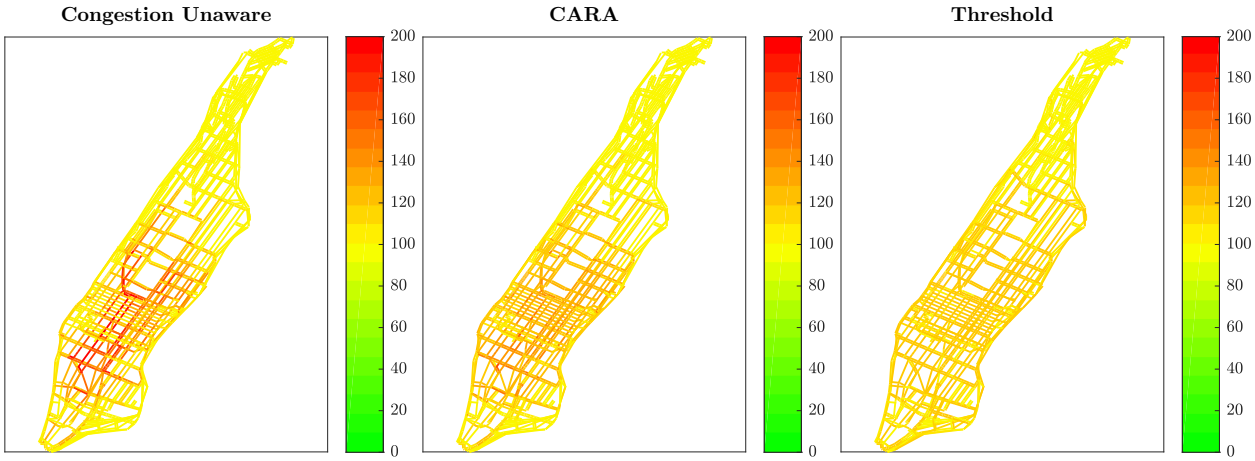

FIGURE 5 Road congestion map for the naïve algorithm, CARA, and the threshold model with a 100% exogenous road usage level.

on main roads and favors longer detours.

On the other hand, CARA improves the operational cost and the travel time compared to both baseline approaches for all levels of exogenous road usage, as the error introduced by the convex relaxation (11) is well bounded below 10%. In dire situations with heavy traffic (when CA algorithms are most needed) CARA achieves improvements of up to 20% in terms of travel time, striking an effective compromise between congesting roads and detouring vehicles, as shown in Fig. 5.

To conclude, our experiments show that CARA can significantly reduce both operational cost and the duration of customer trips. On the other hand, although conservative (thresholded) models of congestion offer a realistic evaluation of travel constraints at low- to mid-levels of exogenous traffic, at high-levels they perform worse than a naïve approach that ignores congestion effects altogether. Most significantly, the overconservative nature of the threshold model leads to an increase of more than $0.50 USD per trip. Conversely, the balanced nature of CARA leads to a decrease of the same amount per trip. With more than half a million trips happening on average each day in NYC (*30*), this efficiency could save over one hundred million dollars over the course of one year.

4. CONCLUSION

In this paper we studied the impact of congestion-aware routing in Autonomous Mobility-on-Demand systems. By leveraging an affine approximation to the Bureau of Public Roads model, we derived a congestion-aware routing algorithm (CARA) that entails solving a computationally-efficient, convex quadratic program. We applied CARA to a case study of AMoD traffic routing in Manhattan and compared its performance to two baselines: a congestion-unaware routing algorithm and an algorithm that considers congestion via a threshold model. Our results show that CARA always outperforms the status quo baselines and achieves an effective balance between congesting roads and detouring vehicles, thus achieving significant improvements, especially under high-levels of traffic. Additionally, we showed that in certain cases threshold models may even worsen the system performance compared to a congestion-unaware approach. Whereas considering congestion as a hard constraint can improve performance under low- to medium-levels of

traffic, it can be overconservative and therefore detrimental in terms of travel time for high-levels of congestion.

This work opens several avenues for future research. Foremost, we would like to design an operational algorithm to apply CARA at a microscopic level. We note that, because the mesoscopic solutions presented here were computed in less than four minutes, the extension to an operational algorithm is completely attainable. Namely, CARA can feasibly become operational by continually solving for optimal flows given updated, real-time customer demands in a receding-horizon model predictive fashion (cf. (*31–33*)). Furthermore, as discussed in Section 2.4, near-optimal integer-valued flows for operational implementation can be achieved from the fractional solutions computed with CARA via random sampling algorithms (*27*). Additional future research includes coupling CARA with urban infrastructure, e.g., the power grid (*23*) and public transit (*22*); including stochastic effects, e.g., of customer demand and road congestion (*34*); and adapting the cost function to include different metrics based on customer preferences.

5. ACKNOWLEDGEMENTS

We would like to thank Dr. Guido Gentile and Dr. Daniele Vigo for providing useful insights on congestion models, and Dr. Ilse New for proofreading this paper and providing us with her expert comments and advice. The first author would like to express his gratitude to Dr. Chris Onder for his support. This research was supported by the National Science Foundation under CAREER Award CMMI-1454737 and the Toyota Research Institute (TRI). This article solely reflects the opinions and conclusions of its authors and not NSF, TRI, or any other entity.

6. STATEMENT OF CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: Mauro Salazar, Marco Pavone; data collection: Mauro Salazar, Matthew Tsao; analysis and interpretation of results: Mauro Salazar, Matthew Tsao, and Maximilian Schiffer; draft manuscript preparation: Mauro Salazar, Matthew Tsao, Izabel Aguiar, Maximilian Schiffer, and Marco Pavone. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

[1] *The World Factbook*. Central Intelligence Agency, 2018, Available at https://www.cia.gov/library/publications/the-world-factbook/fields/2212.html.

[2] Siddiqui, F., *Failing transit ridership poses an 'emergency' for cities, experts fear*. The Washington Post, 2018, available online.

[3] Molla, R., *Americans seem to like ride-sharing services like Uber and Lyft. But it's hard to say exactly how many use them.* Recode, 2018, Available at https://www.recode.net/2018/6/24/17493338/ride-sharing-services-uber-lyft-how-many-people-use.

[4] Bureau of Transportation Statistics, *Transportation Statistics Annual Report 2017*. U.S. Dept. of Transportation, 2017.

[5] Hu, W., *Your Uber Car Creates Congestion. Should You Pay a Fee to Ride?* The New York Times, 2017, available online.

[6] Spieser, K., K. Treleaven, R. Zhang, E. Frazzoli, D. Morton, and M. Pavone, Toward a Systematic Approach to the Design and Evaluation of Autonomous Mobility-on-Demand Systems: A Case Study in Singapore. In *Road Vehicle Automation*, Springer, 2014.

[7] *Live Map*. Waze Mobile, 2018, Available at https://www.waze.com/livemap.

[8] *Google Maps Platform: Routes*. Google, 2018, Available at https://cloud.google.com/maps-platform/routes/.

[9] Bureau of Public Roads, *Traffic Assignment Manual*. U.S. Dept. of Commerce, Urban Planning Division, 1964.

[10] Spiess, H., Technical Note—Conical Volume-Delay Functions. *Transportation Science*, Vol. 24, No. 2, 1990, pp. 153–158.

[11] Rivas, A., G. Inmaculada, S. Sánchez-Cambronero, R. M. Barba, and L. Ruiz-Ripoll, A Continuous Dynamic Traffic Assignment Model From Plate Scanning Technique. *Transport Research Procedia*, Vol. 18, 2016, pp. 332–340.

[12] Aslam, J., S. Lim, and D. Rus, Congestion-aware Traffic Routing System using sensor data. In *Proc. IEEE Int. Conf. on Intelligent Transportation Systems*, 2012.

[13] Manley, E., T. Cheng, A. Penn, and A. Emmonds, A framework for simulating large-scale complex urban traffic dynamics through hybrid agent-based modelling. *Computers, Environment and Urban Systems*, Vol. 44, 2014, pp. 27–36.

[14] Treiber, M., A. Hennecke, and D. Helbing, Microscopic simulation of congested traffic. In *Traffic and Granular Flow '99*, Springer Berlin Heidelberg, 2000.

[15] Yang, Q. and H. N. Koutsopoulos, A microscopic traffic simulator for evaluation of dynamic traffic management systems. *Transportation Research Part C: Emerging Technologies*, Vol. 4, No. 3, 1996, pp. 113–129.

[16] Balmer, M., M. Rieser, K. Meister, D. Charypar, N. Lefebvre, and K. Nagel, MATSim-T: Architecture and Simulation Times. In *Multi-Agent Systems for Traffic and Transportation Engineering*, 2009, chap. 3.

[17] Fagnant, D. J. and K. M. Kockelman, The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. *Transportation Research Part C: Emerging Technologies*, Vol. 40, 2014, pp. 1–13.

[18] Levin, M. W., K. M. Kockelman, S. D. Boyles, and T. Li, A general framework for modeling shared autonomous vehicles with dynamic network-loading and dynamic ride-sharing application. *Computers, Environment and Urban Systems*, Vol. 64, 2017, pp. 373 – 383.

[19] Janson, B. N., Dynamic traffic assignment for urban road networks. *Transportation Research Part B: Methodological*, Vol. 25, No. 2–3, 1991, pp. 143–161.

[20] Peeta, S. and H. S. Mahmassani, System optimal and user equilibrium time-dependent traffic assignment in congested networks. *Annals of Operations Research*, Vol. 60, No. 1, 1995, pp. 81–113.

[21] Rossi, F., R. Zhang, Y. Hindy, and M. Pavone, Routing Autonomous Vehicles in Congested Transportation Networks: Structural Properties and Coordination Algorithms. *Autonomous Robots*, 2018, in Press.

[22] Salazar, M., F. Rossi, M. Schiffer, C. H. Onder, and M. Pavone, On the Interaction between Autonomous Mobility-on-Demand and the Public Transportation Systems. In *Proc. IEEE Int. Conf. on Intelligent Transportation Systems*, 2018, in Press. Extended Version, Available at `https://arxiv.org/abs/1804.11278`.

[23] Rossi, F., R. Iglesias, M. Alizadeh, and M. Pavone, On the Interaction Between Autonomous Mobility-on-Demand Systems and the Power Network: Models and Coordination Algorithms. In *Robotics: Science and Systems*, 2018, Extended version available at `https://arxiv.org/abs/1709.04906`.

[24] Guzzella, L. and A. Sciarretta, *Vehicle Propulsion Systems*. Springer Berlin Heidelberg, 2007.

[25] Neuburger, H., The economics of heavily congested roads. *Transportation Research*, Vol. 5, No. 4, 1971, pp. 283–293.

[26] Wardrop, J. G., Some Theoretical Aspects of Road Traffic Research. *Proc. of the Institution of Civil Engineers*, Vol. 1, No. 3, 1952, pp. 325–362.

[27] Rossi, F., *On the Interaction between Autonomous Mobility-on-Demand Systems and the Built Environment: Models and Large Scale Coordination Algorithms*. Ph.D. thesis, Stanford University, Dept. of Aeronautics and Astronautics, 2018.

[28] Sugar, R., *Uber and Lyft cars now outnumber yellow cabs in NYC 4 to 1*. Curbed, 2017, Available at `https://ny.curbed.com/2017/10/13/16468716/uber-yellow-cab-nyc-surpass-ridership`.

[29] Haklay, M. and P. Weber, OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, Vol. 7, No. 4, 2008, pp. 12–18.

[30] Warerkar, T., *Uber surpasses yellow cabs in average daily ridership in NYC*. Curbed, 2017, Available at `https://ny.curbed.com/2017/10/13/16468716/uber-yellow-cab-nyc-surpass-ridership`.

[31] Seow, K. T., N. H. Dang, and D. H. Lee, A collaborative multiagent taxi-dispatch system. *IEEE Transactions on Automation Sciences and Engineering*, Vol. 7, No. 3, 2010, pp. 607–616.

[32] Pavone, M., S. L. Smith, E. Frazzoli, and D. Rus, Robotic Load Balancing for Mobility-on-Demand Systems. *Int. Journal of Robotics Research*, Vol. 31, No. 7, 2012, pp. 839–854.

[33] Zhang, R., F. Rossi, and M. Pavone, Model Predictive Control of Autonomous Mobility-on-Demand Systems. In *Proc. IEEE Conf. on Robotics and Automation*, 2016.

[34] Tsao, M., R. Iglesias, and M. Pavone, Stochastic Model Predictive Control for Autonomous Mobility on Demand. In *Proc. IEEE Int. Conf. on Intelligent Transportation Systems*, 2018, in Press. Extended Version, Available at `https://arxiv.org/pdf/1804.11074`.