Congestion-aware Routing and Rebalancing of Autonomous Mobility-on-Demand Systems in Mixed Traffic*

Salomón Wollenstein-Betech[1], Arian Houshmand[1], Mauro Salazar[2], Marco Pavone[2], Christos G. Cassandras[1], and Ioannis Ch. Paschalidis[1]

***Abstract*— This paper studies congestion-aware route-planning policies for Autonomous Mobility-on-Demand systems, whereby a fleet of autonomous vehicles provides on-demand mobility under mixed traffic conditions. Specifically, we first devise a convex model to optimize the AMoD routing and rebalancing decisions capturing their endogenous impact on travel time. Second, we include this model in an iterative framework to account for reactive exogenous traffic that selfishly adapts to the AMoD routes by minimizing its travel time in a user-centric fashion. Finally, we showcase the effectiveness of our framework with two case-studies considering the sub-networks of the Eastern Massachusetts and the New York City transportation networks. Our results suggest that for high levels of demand pure AMoD travel can be detrimental due to the additional traffic stemming from the rebalancing flows, whilst the combination of AMoD with walking or micromobility options can significantly improve the overall system performance.**

***Index Terms*— Mobility-on-Demand, System-Centric Routing, Rebalancing, Intermodal Mobility, Mixed Autonomy.**

I. Introduction

In the past decade, the rapid adoption of smartphone technologies and wireless communications coupled with the emergance of sharing economies has resulted in a widespread use of Mobility-on-Demand (MoD) services. One of the main operational challenges that these services face is deciding vehicle routes and rebalancing policies of a fleet. Currently, MoD systems use routing services (e.g., Waze and Google Maps) to route their vehicles, and dynamic pricing plus a real-time heat-map of users' demand to rebalance their fleets.

Given this *user-centric* approach to route vehicles, in which every driver acts selfishly to minimize their own travel time, the network reaches an equilibrium known as the *Wardrop* equilibrium [1]. Unfortunately, these equilibria are in general suboptimal compared to the system optimum, achievable when vehicles are coordinated by a central controller in a *system-centric* fashion.

Recently, the combination of MoD services with Connected and Automated Vehicles (CAVs) has attracted the interest of academia and industry, giving rise to Autonomous Mobility-on-Demand (AMoD) systems (see Fig. 1). These are fleets of CAVs providing on-demand mobility expected to reduce labor costs, accidents, harmful emissions [2], and increase the efficient operation of fleets by uding a central controller. In this context, this paper presents system-optimal routing and rebalancing strategies for AMoD systems in mixed-traffic conditions.

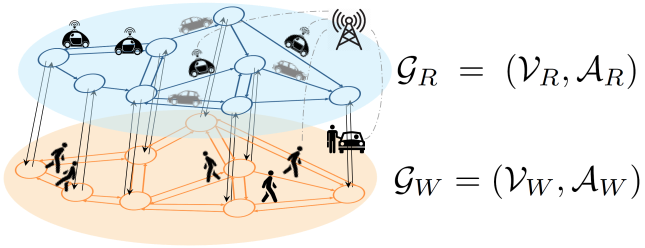


Fig. 1: AMoD network (supergraph) consisting of two digraphs for the road (blue) and the walking (orange) network; the black arrows represent switching arcs. AMoD vehicles are in black and private vehicles in grey.

Related literature: AMoD systems and rebalancing policies have been extensively studied using simulation models [3]–[5], queuing-theoretical models [6], [7], and network-flow models [8], [9]. In [3], the rebalancing of an AMoD system is addressed using a data-driven real-time parametric controller. Alternatively, in [8], the rebalancing problem is studied using a steady-state fluid model. Although these models seek to find a good rebalancing policy, they do not consider the routing decisions for AMoD users when Private Vehicles are present in the network.

Few work has been done to solve this problem jointly (routing and rebalancing), and most of it uses threshold approximations of the travel time function. In [9], the authors show that under relatively mild assumptions rebalancing vehicles do not lead to an increase in congestion, this suggests that the joint problem can be decoupled without having an important impact in the solution. Moreover, [10] introduced a piecewise affine approximation of the travel time function in order to relax the non-convex joint problem to a quadratic program. Yet, depending on the congestion levels, the latter approach may lack in accuracy. Moreover, [9] and [10] assume non-reactive exogenous traffic flows. Recently, reactive private traffic was modeled in [11] to show that under a system-centric optimal-routing strategy both CAVs and non-CAVs can achieve better performance in terms of travel time and energy savings. However, such an approach neither captures rebalancing effects nor intermodal routing possibilities.

Statement of contribution: This paper bridges the gap

*This work was supported in part by NSF under grants ECCS-1509084, DMS-1664644, CNS-1645681, IIS-1914792, and CMMI-1454737, by AFOSR under grant FA9550-19-1-0158, by ARPA-E's NEXTCAR program under grant DEAR0000796, by the MathWorks, by the ONR under grant N00014-19-1-2571, by the NIH under grant 1R01GM135930, and by the Toyota Research Institute (TRI). This article solely reflects the opinions and conclusions of its authors and not NSF, TRI, or any other entity. The authors thank D. Sverdlin-Lisker and Dr. I. New for proofreading this paper.

[1] The authors are with the Division of Systems Engineering, Boston University, Brookline, MA 02446 USA {salomonw, arianhm cgc, yannisp}@bu.edu

[2] The authors are with the department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94325 USA {samauro,pavone}@stanford.edu

between [10] and [11]. Specifically, we study how system-optimal routing of AMoD services can affect the system-level performance in *mixed-traffic* (presence of AMoDs and private vehicles in the road network). Similar to [10] we assume that AMoD users can use multiple modes of transportation, i.e., autonomous taxi rides and walking. In addition to [9], [10], we let the private vehicle flow to be *reactive*, meaning that private vehicles will choose their routes selfishly considering the AMoD flow. To this end, we use the framework previously developed in [11] for modeling the interaction between AMoDs and private vehicles. Moreover, we devise a more accurate approximation of the travel time function, while still maintaining the convex structure of the AMoD problem. The proposed model can efficiently compute the congestion-aware routing and the rebalancing policy for a given demand and road network topology. Finally, with this framework at hand, we analyze the trade-offs between the benefit of socially-routing and the cost of rebalancing. Moreover, we provide examples of the advantages of providing mobility services by considering autonomous taxi-rides, walking and micromobility options.

Organization: The rest of the paper is organized as follows: In Section II we provide preliminaries of the model and its formulation. In Section III we develop a convex approximation of the original problem to overcome its non-convex nature. We present experiments using the Eastern Massachusetts and New York City road networks in Section IV. Finally, in Section V we conclude the paper and point to future research directions.

Notation: All vectors are column vectors and denoted by bold lowercase letters. Bold uppercase letters denote matrices. We use "prime" to denote the transpose of a matrix or vector. We denote by $\mathbb{1}_{\mathbb{x}}$ the indicator function. $\|\cdot\|$ denotes the ℓ_2 norm.

II. Problem Formulation

In this section, we present mesoscopic models for planning the routes and rebalancing strategies used throughout the paper. First, we introduced the notation and preliminaries of transportation modeling. With this in hand, we model the system-centric routing and rebalancing of AMoD followed by the user-centric model assumend for Private Vehicles. Finally, we formulate the joint problem of congestion-aware routing and rebalancing of AMoD in mixed traffic.

A. Preliminaries

Consider an AMoD system which provides mobility services through two modes of transportation: walking and autonomous taxi-rides. To model the system, let $\mathcal{G}$ be a network (supergraph) composed of two layers, a road and a walking network. We denote by $\mathcal{G}_{\mathrm{R}} = (\mathcal{V}_{\mathrm{R}}, \mathcal{A}_{\mathrm{R}})$ the road network and by $\mathcal{G}_{\mathrm{W}} = (\mathcal{V}_{\mathrm{W}}, \mathcal{A}_{\mathrm{W}})$ the pedestrian graph where $(\mathcal{V}_{\mathrm{R}}, \mathcal{A}_{\mathrm{R}})$ and $(\mathcal{V}_{\mathrm{W}}, \mathcal{A}_{\mathrm{W}})$ are the sets of intersections (vertices) and streets (arcs) in the road and in the pedestrian network, respectively. Then, the supergraph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ is composed of $\mathcal{G}_{\mathrm{R}}$ and $\mathcal{G}_{\mathrm{W}}$, and a set of *switching* arcs $\mathcal{A}_{\mathrm{S}} \subset \mathcal{V}_{\mathrm{R}} \times \mathcal{V}_{\mathrm{W}} \cup \mathcal{V}_{\mathrm{W}} \times \mathcal{V}_{\mathrm{R}}$ that connect the pedestrian and the road network layers to allow AMoD users to change modes (see Fig. 1). Formally $\mathcal{G}$ is composed of the set of vertices $\mathcal{V} = \mathcal{V}_{\mathrm{R}} \cup \mathcal{V}_{\mathrm{W}}$ and arcs $\mathcal{A} = \mathcal{A}_{\mathrm{R}} \cup \mathcal{A}_{\mathrm{W}} \cup \mathcal{A}_{\mathrm{S}}$.

In order to model the demanded trips, let $\mathbf{w} = (w_s, w_t)$ denote an Origin-Destination (OD) pair and $d_{\mathbf{w}} \geq 0$ the demand rate at which customers request service per unit time from origin w_s to destination w_t. Let W be the total number of OD pairs and $\mathcal{W} = \{\mathbf{w}_k : \mathbf{w}_k = (w_{sk}, w_{tk}), k = \{1, ..., W\}\}$ the set of OD pairs. Let us denote a vectorized version of the demand $\mathbf{g} = (d^{\mathbf{w}} : \mathbf{w} \in \mathcal{W})$ denoting the demand flows for all OD pairs.

To keep track of AMoD users' flow on an arc, we let $x_{ij}^{\mathbf{w}}$ denote the AMoD flow induced by OD pair $\mathbf{w}$ in link $(i,j) \in \mathcal{A}$. Given that the AMoD needs to rebalance its vehicles to ensure service, we let x_{ij}^r be the *rebalancing flow* on road (i,j). Finally, to consider the interaction between the AMoD provider and the other vehicles, we let x_{ij}^p be the self-interested (*private vehicle*) flow on (i,j). We use the term *private* as we assume that self-interested users must arrive at their destination with their vehicle and do not have the option of switching transportation mode (i.e., walking). To simplify notation, we call the AMoD user flow of vehicles on a road x_{ij}^u, which is equal to

$$x_{ij}^u = \sum_{\mathbf{w} \in \mathcal{W}} x_{ij}^{\mathbf{w}}, \tag{1}$$

and the total flow on a link of G_{R} be

$$x_{ij} = x_{ij}^u + x_{ij}^r + x_{ij}^p. \tag{2}$$

Let $t_{ij}(x) : \mathbb{R}_+^{|\mathcal{A}_{\mathrm{R}}|} \mapsto \mathbb{R}_+$ be the *travel time* function, i.e., the time it takes to cross link (i,j) given the flow on the link. Using the same function structure as in [12], we characterize t_{ij} as a function of the flow x_{ij} with

$$t_{ij}(x_{ij}) = t_{ij}^0 f(x_{ij}/m_{ij}), \tag{3}$$

where m_{ij} is the road's capacity, $f(\cdot)$ is a strictly increasing, positive, and continuously differentiable function, and t_{ij}^0 is the free-flow travel time on link (i,j). We would like to consider functions with $f(0) = 1$, which ensures that if there is no flow on the link, the travel time t_{ij} is equal to the free-flow travel time. Typically, travel time functions used by urban planners and researchers are polynomials which are hard to estimate [13]. A widely used function is the *Bureau of Public Roads (BPR)* travel time function [14] denoted by

$$t_{ij}(x_{ij}) = t_{ij}^0(1 + 0.15(x_{ij}/m_{ij})^4). \tag{4}$$

Throughout this paper, we use this function to decide the routes of AMoD users and private vehicles, given the network flow levels. For AMoD users who walk, we consider a constant travel time (independent of the flow) on each link.

B. System-centric Routing and Rebalancing of AMoD

Recall that our goal is to find the system-centric congestion-aware routes and rebalancing policy of an AMoD provider. The objective consists of minimizing the cost composed of the overall travel time of AMoD users, and a regularizer penalizing rebalancing flow.

We formulate the problem similar to [10] where we address the problem from an AMoD perspective. Let $d_{\mathbf{w}}^u$ be customer requests to the AMoD provider traveling from

origin w_s to destination w_t. The problem we aim to solve is then expressed by

$$\min_{\{\mathbf{x}^{\mathbf{w}}\}_{\mathbf{w}\in\mathcal{W}},\mathbf{x}^r} J(\mathbf{x}) := \sum_{(i,j)\in\mathcal{A}} t_{ij}(x_{ij})x^u_{ij} + \mathbf{c}'\mathbf{x}^r \tag{5a}$$

$$\text{s.t.} \sum_{i:(i,j)\in\mathcal{A}} x^{\mathbf{w}}_{ij} + \mathbb{1}_{j=w_s} d^u_{\mathbf{w}} = \sum_{k:(j,k)\in\mathcal{A}} x^{\mathbf{w}}_{jk} + \mathbb{1}_{j=w_t} d^u_{\mathbf{w}}, \quad \forall \mathbf{w}\in\mathcal{W}, j\in\mathcal{V}, \tag{5b}$$

$$\sum_{i:(i,j)\in\mathcal{A}_{\mathrm{R}}} \left(x^r_{ij} + x^u_{ij}\right) = \sum_{k:(j,k)\in\mathcal{A}_{\mathrm{R}}} \left(x^r_{jk} + x^u_{jk}\right), \quad \forall j\in\mathcal{V}_{\mathrm{R}}, \tag{5c}$$

$$\mathbf{x}^{\mathbf{w}}_{ij} \geq 0, \quad \forall \mathbf{w}\in\mathcal{W}, (i,j)\in\mathcal{A}, \tag{5d}$$

$$\mathbf{x}^r_{ij} \geq 0, \quad \forall \mathbf{w}\in\mathcal{W}, (i,j)\in\mathcal{A}_{\mathrm{R}}, \tag{5e}$$

where we use bold notation $\mathbf{x}$ to represent a vector containing all the elements of x_{ij}. Moreover, constraints (5b) takes care of flow conservation and demand compliance as in a multi-commodity transportation problem, constraints (5c) ensure the rebalancing of the AMoD fleet, and the last two sets of constraints (5d)-(5e) restrict the flows to non-negative values.

The objective J is composed of two terms. The first term considers the total travel time of AMoD users. This term evaluates the travel time function $t_{ij}(x_{ij})$ with respect to the total flow (see (2)) which includes variables corresponding to private vehicle flow x^p_{ij} (assumed to be fixed), and the rebalancing flow x^r_{ij}. Hence, when taking the product of $t_{ij}(x_{ij})x^u_{ij}$ we obtain a non-convex function. To address the non-convexity issue, we use a piecewise affine approximation of $t_{ij}(x_{ij})$ which is further presented in Section III.

The second term, i.e., $\mathbf{c}'\mathbf{x}^r$, acts as a linear reguralizer whose purpose is to penalize rebalancing flows. This will ensure that a cost for rebalancing of the fleet is taken into account. In this work, we use $\mathbf{c} = \lambda\mathbf{t}^0$. One can think of this reguralizer as a linear travel time function with respect to the rebalancing flow (since $(\lambda\mathbf{t}^0)'\mathbf{x}^r$). Hence, if one lets λ be high, with respect to the overall travel time, the reguralizer term will dominate the objective. Hence, we use a small λ in order to guide the rebalancing flow through good paths without dominating the AMoD user routing decisions.

C. Private Vehicle Flow Modeling

Aiming to understand the interaction between a system-centric AMoD fleet and self-interested private vehicles, we assume some rationale behind private vehicle decisions. To model this class of vehicles we use the *user-centric* approach as in the Traffic Assignment Problem (TAP) [15]. This model finds, given OD demands, the flows in the network which achieve a Wardrop equilibrium [1].

Given a demand $\mathbf{g}^p$ for this type of vehicle, each private user decides its route such that it minimizes its own travel time. Moreover, we impose that private vehicles can travel exclusively through the road network $\mathcal{G}_{\mathrm{R}}$. In other words, we do not allow private vehicles to change their transportation mode to walking.

Let $x^{p,\mathbf{w}}_{ij}$ be the flow in link (i,j) induced by private vehicle demand $d^p_{\mathbf{w}}$ of OD pair $\mathbf{w}$. Then, we assume private vehicles to decide their routes by using the user-centric approach, we modeled them using

$$\min_{\mathbf{x}^p} \sum_{(i,j)\in\mathcal{A}_{\mathrm{R}}} \int_{x^u_{ij}+x^r_{ij}}^{x_{ij}} t_{ij}(s)ds \tag{6a}$$

$$\text{s.t} \sum_{i:(i,j)\in\mathcal{A}_{\mathrm{R}}} x^{p,\mathbf{w}}_{ij} + d^p_{\mathbf{w}}\mathbb{1}_{j=w_s} = \sum_{k:(j,k)\in\mathcal{A}_{\mathrm{R}}} x^{p,\mathbf{w}}_{jk} + d^p_{\mathbf{w}}\mathbb{1}_{j=w_t}, \quad \forall \mathbf{w}\in\mathcal{W}, j\in\mathcal{V}_{\mathrm{R}}, \tag{6b}$$

$$\mathbf{x}^{p,\mathbf{w}} \geq \mathbf{0}. \tag{6c}$$

Notice that this version of the user-centric TAP is slightly different to the typical one, given that it considers the AMoD flow in its objective (see limits of the integral on (6a)).

To solve this problem we assume that the AMoD flow is fixed and private vehicles plan their routes considering AMoD flows as exogenous. When using this restriction, we can use the *Method of Successive Averages* (MSA) [16] to solve (6). Let us use the shorthand notation of $\texttt{TAP}(\mathbf{g}, \mathbf{x}^e)$ to indicate the TAP with $\mathbf{x}^e$ being the exogenous flow in the network. We denote a solution to the TAP problem (6) by $\mathbf{x}^p = \min \texttt{TAP}(\mathbf{g}^p, \mathbf{x}^u + \mathbf{x}^r)$.

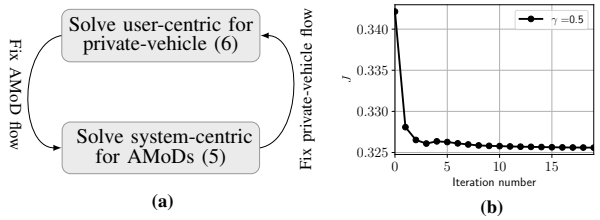
D. System-centric Routing and Rebalancing of AMoD in Mixed Traffic

In this section we introduce an iterative approach to find the optimal system-centric routes (AMoD users) and rebalancing policy in the presence of mixed traffic (AMoD and private flows). We assume the AMoD central controller desires to minimize the time of its fleet as opposed to minimizing the overall travel time for all vehicles in the network. Interestingly, AMoD flows react to the decisions made by private vehicles and these, in turn, react to private vehicles' flows. Hence, whenever privately vehicles make their routing decisions, the AMoD fleet adjusts theirs, and vice versa. This creates a nested optimization problem between these two classes of vehicles. To give a formal definition of this game-theoretic problem we use the following bi-level optimization framework

$$\min_{\{\mathbf{x}^{\mathbf{w}}\}_{\mathbf{w}\in\mathcal{W}},\mathbf{x}^r,\mathbf{x}^p} J(\mathbf{x})$$
$$\text{s.t.} \quad \text{(5b)} - \text{(5e)},$$
$$\mathbf{x}^p \in \arg\min \texttt{TAP}(\mathbf{g}^p, \mathbf{x}^u + \mathbf{x}^r), \tag{7a}$$

which has the same structure as (5) with the additional constraint (7a). This constraint refers to the TAP (the lower-level problem), which depends on the solution of the full problem (upper-level). Note that the upper-level problem is minimizing over the AMoD users, rebalancing, and privately-owned vehicle flows.

This phenomenon has been identified and is often described in a *Stackelberg game* framework. In this setting, there is a *leader* agent (in our case the AMoD manager) and a *follower* (private vehicles). In transportation networks, Korilis et al. [17] derived sufficient conditions to solve this problem when the network has parallel links. More recently, Lazar et al. [18] have analyzed the links' capacity and price of anarchy for mixed traffic given parallel links. Although these models enable a better understanding of the



(a) (b)

Fig. 2: (a): A sketch of the procedure for solving the bi-level problem (7). (b): An example of the total cost converging for an AMoD penetration rate of 0.5 on the NYC sub-network

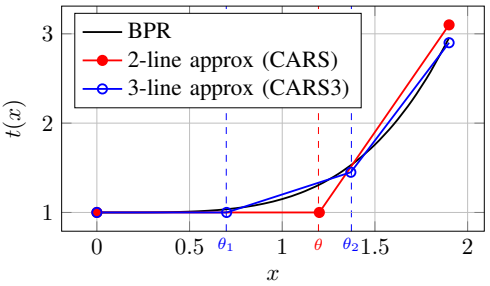


Fig. 3: Travel time function approximation

phenomenon, they are not applicable to general networks and one can hardly assess the benefits of system-centric routing in real-world networks. To address this limitation, we use the framework developed by Houshmand et al. [11], which uses an iterative approach to reach an equilibrium between the private and AMoD flows (Fig. 2).

Instead of solving the bi-level Problem (7), we solve Problems (5) and (6) iteratively and use the output of each problem as the input to the other one. In other words, consider a private vehicle demand $\mathbf{g}^u$ and solve $\mathbf{x}^p = \min \texttt{TAP}(\mathbf{g}^p, \mathbf{0})$. Then, solve the AMoD routing and rebalancing problem (5) for AMoD demand $\mathbf{g}^u$ with fixed input $\mathbf{x}^p$ (the solution of the previously solved TAP). Since private vehicles were unaware of AMoDs in the system while solving the TAP, we solve again the problem considering a fixed flow equal to $\mathbf{x}^u + \mathbf{x}^r$, i.e., $\mathbf{x}^p = \texttt{TAP}(\mathbf{g}^p, \mathbf{x}^u + \mathbf{x}^r)$, and iterate this process until it converges as shown in Fig. 2b.

III. AMoD Routing and Rebalancing Problem

As mentioned earlier, the problem of routing and rebalancing as stated in (5) is non-convex for typical travel time functions such as BPR. This happens due to the term $t(x_{ij})x^r_{ij}$ in the objective function which take products of the form $k(x^u_{ij})^n x^r_{ij}$ with k and n being a constant and the order of the polynomial, respectively. To overcome this issue, we take the suggested piecewise affine approximation proposed in [10] and extend it to a 3-lines approximation. Specifically, we first present the analysis for the 2-lines Congestion-Aware Routing Scheme (CARS) [10] and then extend it to the 3-line segment case (CARS3). Finally, we present a disjoint formulation of the problem which will serve as a benchmark for comparison.

A. 2-lines Piecewise Affine Approximation (CARS)

Recall that the non-convexity in (5a) arises from the product of the AMoD users flow x^u_{ij} with the rebalancing flow x^r_{ij}. Hence, we aim to approximate this term with a convex function which makes it more computationally efficient, and therefore gives tractability to larger instances of the problem. Specifically, we approximate the latency function (Eq. (3)) using a piecewise affine function as shown in Fig. 3. Let such a function be

$$\hat{t}_{ij}(x) = \begin{cases} at^0_{ij}, & \text{if } x < \theta_{ij}, \\ at^0_{ij} + b_{ij}(\theta_{ij} - x), & \text{if } x \geq \theta_{ij}, \end{cases} \tag{8}$$

where a and b_{ij} are constant values. In our case, we assume $a = 1$, and $b_{ij} = \beta / m_{ij}$ with β being the slope of the second segment. Let the non-smooth threshold of the function be $\theta_{ij} = m_{ij}\theta$, where θ is the threshold in the normalized travel time function. In order to model this non-smooth function in the optimization problem, we introduce the set of slack variables ε_{ij} defined as

$$\varepsilon_{ij} = \max\{0, x_{ij} - \theta_{ij}\}, \tag{9}$$

which denotes the exceeding flow after threshold θ_{ij}. In the optimization problem (5) we model these variables by adding linear constraints $\varepsilon_{ij} \geq 0$ and $\varepsilon_{ij} \geq \theta_{ij} - x$, provided that the objective is an increasing function of ε_{ij}. With these definitions we are ready to analyze and propose a tractable cost function. To that end , we focus attention on an element-wise analysis of the first term (non-convex part) of objective (5a) using $\hat{t}$ instead of t, which we call $\hat{J}_{ij}$.

$$\begin{aligned} \hat{J}_{ij} &= \hat{t}_{ij}(x_{ij})x^u_{ij} && \text{(10a)} \\ &= t^0_{ij}(a + b_{ij}\varepsilon_{ij})x^u_{ij} && \text{(10b)} \\ &= at^0_{ij}x^u_{ij} + t^0_{ij}b_{ij}\varepsilon_{ij}x^u_{ij} && \text{(10c)} \\ &= at^0_{ij}x^u_{ij} + b_{ij}t^0_{ij}\varepsilon_{ij}(\varepsilon_{ij} + \theta_{ij} - x^r_{ij} - x^p_{ij}) && \text{(10d)} \\ &\leq at^0_{ij}x^u_{ij} + b_{ij}t^0_{ij}\varepsilon_{ij}(\varepsilon_{ij} + \theta_{ij} - x^p_{ij}) && \text{(10e)} \end{aligned}$$

where in (10d) we express x^u_{ij} by using a combination of (2) and (9); in the last step (10e), we add to $\hat{J}_{ij}$ the term $b_{ij}t^0_{ij}\varepsilon_{ij}x^r_{ij}$. By adding this term to $\hat{J}$, we consider a relaxation of the original problem (i.e., minimizing an upper bound of $\hat{J}$). This relaxation allows the proposed objective to be quadratic. Let the relaxed objective be

$$\begin{aligned} J^{\text{QP}}_{ij} &= at^0_{ij}x^u_{ij} + b_{ij}t^0_{ij}\varepsilon_{ij}(\varepsilon_{ij} + \theta_{ij} - x^p_{ij}) && \text{(11a)} \\ &= \hat{t}_{ij}(x_{ij})x^u_{ij} + \hat{t}^{a=0}_{ij}(x_{ij})x^r_{ij}. && \text{(11b)} \end{aligned}$$

where $\hat{t}^{a=0}(x)$ is equal to $\hat{t}(x)$ with parameter $a = 0$. By analyzing this convex approximation J^{QP} with both J and $\hat{J}$, we observe that the implication of adding the extra term is taking into account congestion-aware rebalancing when the flow is greater than θ_{ij}. Nevertheless, this congestion-aware routing of the rebalancing vehicles has a lower impact in J^{QP} than the AMoD users flows since $a = 0$ in $\hat{t}^{a=0}_{ij}(x_{ij})x^r_{ij}$, i.e., the function starts to increase from an initial point equal to zero instead of t^0_{ij}.

Considering that the number of rebalancing vehicles has a minor impact on J in comparison to road congestion, and the fact that it converges to zero for perfectly symmetric demand distributions [9], J^{QP} can be used as a model to estimate the total travel time on road arcs. Our empirical studies shows that, when no rebalancing is considered, the difference between the solution J^* and J evaluated with the

optimal solution of the Quadratic Program (QP) is typically less than 5% (Figs. 5a and 5b).

B. 3-lines Piecewise Affine Approximation (CARS3)

Given that CARS might not provide a very accurate estimate of travel times when the flow is around the capacity level (Fig. 3), we next to approximate the travel time function using a more accurate 3-lines piecewise affine function. To construct this function, we will follow the same analysis as in the 2-lines case. The price to pay for increasing the precision of the function is that it requires adding $|\mathcal{A}|$ (number of arcs) extra variables and $|\mathcal{A}|$ new linear constraints to the optimization problem. Following the same analysis as in the previous section we define

$$\hat{t}_{ij}(x) = \begin{cases} at^0_{ij}, & \text{if } x < \theta_{ij} \\ at^0_{ij}+b_{ij}(\theta^{(1)}_{ij}-x), & \text{if } \theta^{(1)}_{ij} \le x \le \theta^{(2)}_{ij} \\ at^0_{ij}+b_{ij}(\theta^{(2)}_{ij}-\theta^{(1)}_{ij})+c_{ij}(\theta^{(2)}_{ij}-x), & \text{if } \theta^{(2)}_{ij} \le x, \end{cases}$$

where a, b_{ij} and c_{ij} are constant values with $a = 1$, $b_{ij} = \beta/m_{ij}$ and $c_{ij} = \sigma/m_{ij}$. The slope of the function is β for $x_{ij} \in (\theta^{(1)}_{ij}, \theta^{(2)}_{ij})$ and σ for $x_{ij} > \theta^{(2)}_{ij}$. Moreover, $\theta^{(1)}$ and $\theta^{(2)}$ are the normalized, non-smooth thresholds of the travel time function. Assuming $\theta^{(2)}_{ij} \ge \theta^{(1)}_{ij}$ and $\sigma, \beta > 0$ we define two new sets of slack variables as

$$\varepsilon^{(1)}_{ij} = \max\{0, x_{ij} - \theta^{(1)}_{ij} - \varepsilon^{(2)}_{ij}\} \tag{13a}$$
$$\varepsilon^{(2)}_{ij} = \max\{0, x_{ij} - \theta^{(2)}_{ij}\}, \tag{13b}$$

where $\varepsilon^{(1)}_{ij}$ is the excess flow after $\theta^{(1)}_{ij}$ and up to $\theta^{(2)}_{ij} - \theta^{(1)}_{ij}$, and $\varepsilon^{(2)}_{ij}$ is the excess flow after $\theta^{(2)}_{ij}$. Note that $\varepsilon^{(1)}_{ij}$ is defined in terms of $\varepsilon^{(2)}_{ij}$ to ensure that it is upper-bounded by $\theta^{(2)}_{ij} - \theta^{(1)}_{ij}$. Using the same analysis as in the 2-lines case, we get

$$\begin{aligned} \hat{J}_{ij} &= \hat{t}_{ij}(x_{ij})x^u_{ij} && \text{(14a)} \\ &= at^0_{ij}x^u_{ij} + b_{ij}t^0_{ij}\varepsilon^{(1)}_{ij}(\varepsilon^{(1)}_{ij} + \varepsilon^{(2)}_{ij} + \theta^{(1)}_{ij} - x^r_{ij} - x^e_{ij}) \\ &\quad + c_{ij}t^0_{ij}\varepsilon^{(2)}_{ij}(\varepsilon^{(2)}_{ij} + \theta^{(2)}_{ij} - x^r_{ij} - x^e_{ij}) && \text{(14b)} \\ &\le at^0_{ij}x^u_{ij} + b_{ij}t^0_{ij}\varepsilon^{(1)}_{ij}(\varepsilon^{(1)}_{ij} + \varepsilon^{(2)}_{ij} + \theta^{(1)}_{ij} - x^e_{ij}) \\ &\quad + c_{ij}t^0_{ij}\varepsilon^{(2)}_{ij}(\varepsilon^{(2)}_{ij} + \theta^{(2)}_{ij} - x^e_{ij}). && \text{(14c)} \end{aligned}$$

We add the rebalancing variables as in the CARS to get (14c). Even though the term $b_{ij}t^0_{ij}\varepsilon^{(1)}_{ij}\varepsilon^{(2)}_{ij}$ is not guaranteed to be convex, $\varepsilon^{(1)}_{ij}\varepsilon^{(2)}_{ij} = 0$ if $x_{ij} < \theta^{(2)}_{ij}$. Additionally, notice that when $x_{ij} > \theta^{(2)}_{ij}$ the residual flow $\varepsilon^{(1)}_{ij} = (\theta^{(2)}_{ij} - \theta^{(1)}_{ij})$. Therefore, we can replace $b_{ij}t^0_{ij}\varepsilon^{(1)}_{ij}\varepsilon^{(2)}_{ij}$ with $b_{ij}t^0_{ij}(\theta^{(2)}_{ij} - \theta^{(1)}_{ij})\varepsilon^{(2)}_{ij}$ and write the objective function of the QP as

$$\begin{aligned} J^{\text{QP}}_{ij} &= at^0_{ij}x^u_{ij} + b_{ij}t^0_{ij}\varepsilon^{(1)}_{ij}(\varepsilon^{(1)}_{ij} + \theta^{(1)}_{ij} - x^e_{ij}) \\ &\quad + c_{ij}t^0_{ij}\varepsilon^{(2)}_{ij}(\varepsilon^{(2)}_{ij} + \theta^{(2)}_{ij} - x^e_{ij}) \\ &\quad + b_{ij}t^0_{ij}(\theta^{(2)}_{ij} - \theta^{(1)}_{ij})\varepsilon^{(2)}_{ij} \\ &= \hat{t}_{ij}(x_{ij})x^u_{ij} + \hat{t}^{a=0}_{ij}(x_{ij})x^r_{ij}, \end{aligned} \tag{15}$$

where $\varepsilon^{(1)}_{ij}$ and $\varepsilon^{(2)}_{ij}$ are linearly constrained as follows

$$\varepsilon^{(1)}_{ij} \ge 0, \quad \varepsilon^{(1)}_{ij} \ge x_{ij} - \theta^{(1)}_{ij} - \varepsilon^{(2)}_{ij}, \tag{16}$$
$$\varepsilon^{(2)}_{ij} \ge 0, \quad \varepsilon^{(2)}_{ij} \ge x_{ij} - \theta^{(2)}_{ij}. \tag{17}$$

As a result, we get a better convex approximation of the original problem compared to CARS model. The Quadratic Programming (QP) problem is then to minimize (15) subject to (5b)-(5e), and (16)-(17).

An important trade-off worth noting is the difference between CARS and CARS3. Even though CARS3 provides a better approximation of the cost function and hence a better solution to the problem, it requires $|\mathcal{A}|$ additional variables and constraints.

C. Disjoint Strategy

Another way of addressing the system-centric routing and re-balancing problem is to solve the problem using a disjoint method instead of the jointly optimized approach. That is, to solve the system-centric problem for AMoD users first, and then solve the rebalancing problem formulated as a linear program (LP). A formal definition of this problem is first solving

$$\min_{\{\mathbf{x}^w\}_{w\in\mathcal{W}}} \sum_{(i,j)\in\mathcal{A}} t_{ij}(x_{ij})x^u_{ij}, \quad \text{s.t.} \quad \text{(5b)}, \text{(5d)}, \tag{18}$$

and then, using the resulting optimal $\mathbf{x}^{c*}$ to solve

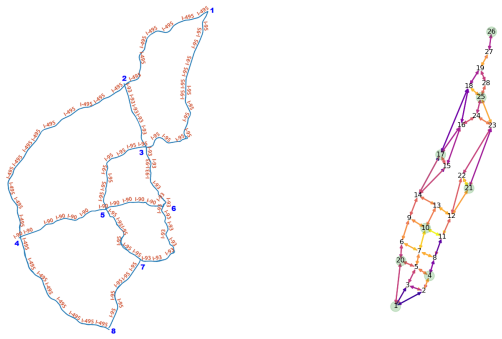
$$\min_{\mathbf{x}^r} \quad \mathbf{c}'\mathbf{x}^r, \quad \text{s.t.} \quad \text{(5c)}, \text{(5e)}, \tag{19}$$

It is important to point out that the system-centric problem (18) is a constrained nonlinear program (NLP) which might take time to solve. In contrast to the disjoint formulation, the methodology we propose (CARS3) offers the possibility to solve the problem as a QP, which is usually faster than a higher order NLP.

It is worth pointing out that both the disjoint problem, and the iterative model proposed in Sec. II-D allow for updating the component $\mathbf{t}^0$ in $\mathbf{c}$ for the travel times $\mathbf{t}(\mathbf{x})$ from the solution of (18) or previous iteration for the dijsoint and iterative method, respectively. This results on a more accurate cost function in terms of the travel time weight for the rebalancing problem.

D. Discussion

A few comments are in order. First, we assume the demand to be time-invariant. This assumption is in order for densely populated urban environments, where requests change slower compared to the average duration of a trip. Second, we use the BPR function to relate traffic flows to travel time and allow flows to be fractional. While not capturing microscopic traffic phenomena, these approximations stem from established modeling assumptions suiting the mesoscopic perspective of our study. Moreover, similar as in [10], our algorithm can be extended to operate in real-time by periodically optimizing the routes with updated information and by recovering integer routes via randomized routing algorithms [19]. Finally, we do not provide theoretical arguments on the uniqueness or stability of the equilibria, due to the non-separability of the cost functions



(a) EMA subnetwork **(b)** NYC subnetwork

Fig. 4: Subnetwork used for the experiments. Details on Appendix B.

with respect to their individual players' strategies [20]. Yet, empirically, the iterative algorithm proposed in Section II-D always converged in few iterations (e.g., see Fig. 2b) to results that are consistent for different penetration rates. We leave the theoretical study of the properties of the equilibria found to future work.

IV. Experiments

In order to validate our proposed routing algorithms, we consider two data-driven case studies on sub-networks of Eastern Massachusetts (EMA) interstate highways and New York City (NYC). The EMA road network (Fig. 4a) consists of 8 nodes and 24 links. We consider every node as a zone (origin-destination candidate) which results in 56 OD pairs. The NYC network was built using two data sources: OpenStreetMaps [21] from which we retrieve the network topology and road characteristics, and the recently released *Uber Movement Speed Data set* [22] which was used to assign speed data to road segments (available hourly). We build sub-network (Fig. 4b) consisting of 28 nodes, 90 edges and 8 Szones (green dots).

We use the three methodologies described in Sec. II-D (CARS, CARS3 and Disjoint) to solve the fleet routing and rebalancing problem and compare their results against each another. Our first two experiments reveal that using CARS and CARS3 result in accurate solutions with low running times for these networks.

A. Accuracy of CARS and CARS3

Using numerical examples, we show how the optimal solution of CARS and CARS3 compare with the optimal solution of the original system-centric problem. To achieve this, we consider the case in which rebalancing is not required, i.e., constraints (5c) are excluded and variables $\mathbf{x}^r$ are set to zero. Then, the non-rebalanced routing problem becomes the system-centric traffic assignment problem with exogenous flow (problem (18)). This problem is convex [15] and can be solved using nonlinear programming (NLP) algorithms.

This experiment assesses the offset of the total cost between the approximate models (CARS, CARS3) and the optimal solution considering the non-rebalancing system-centric model. To make a fair comparison, the solution of CARS and CARS3 are evaluated in the original cost function $J(\mathbf{x})$ from (5a). We gather results for different traffic levels (demands) for the EMA (Fig. 5a) and NYC

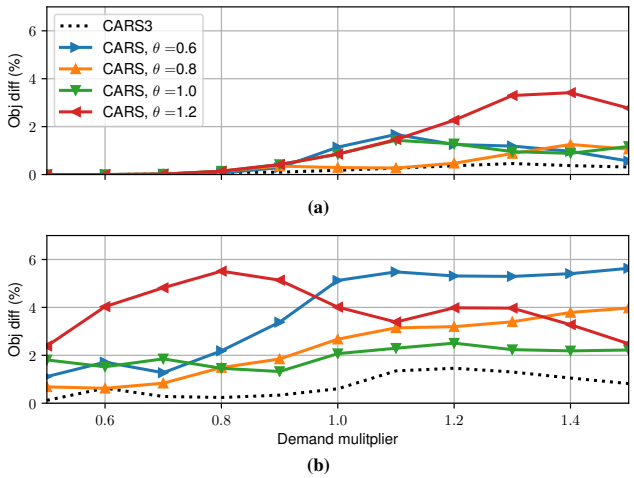


Fig. 5: Deviation in percentage terms between the approximated model and the optimal solution of the non-rebalanced system-centric problem. **(a)** and **(b)** present results for the EMA and NYC networks, respectively.

Model	Type	EMA			NYC		
		μ_τ [s]	σ_τ [s]	$\bar{J}$	μ_τ [s]	σ_τ [s]	$\bar{J}$
CARS	QP	**0.016**	3e-7	0.427	**0.170**	8e-4	0.324
CARS3	QP	**0.022**	1e-6	0.422	**0.215**	3e-3	0.317
Disjoint		**5.48**	0.269	0.421	**24.88**	1.72	0.31
System-centric	NLP	5.48	0.269		24.88	1.72	
Rebalance	LP	2e-4	2e-10		4e-5	2e-10	

TABLE I: Computational times and objective function for different models and networks. μ_τ and σ_τ are the average computational time (seconds) and variance over 30 samples, respectively. The average cost is denoted with $\bar{J}$.

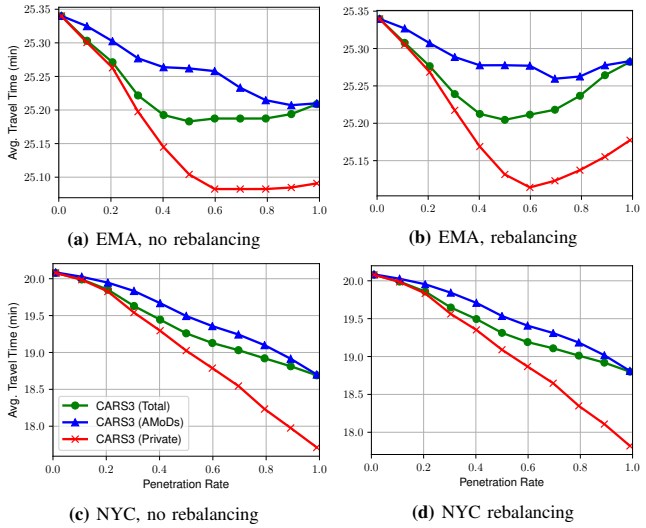
(Fig. 5b) networks. The purpose of using different demands is to investigate the approximation quality of $\hat{t}(\cdot)$ (Fig. 3) at different flow levels. Note that for the two networks, the CARS3 model outperforms the CARS method for different parameters of θ and demand rates. We attribute this behavior to the fact that the 3-lines model yields a more precise approximation to the travel time function than the 2-lines one. Moreover, we consider the CARS3 model to be a good approximation as its deviations are always below 2%.

B. Computational Time and Evaluation of the Cost

We would like to compare the running times of CARS, CARS3, and Disjoint as well as their solutions. For each approach, we solve 30 different problems by multiplying the OD demand vector $\mathbf{g}$ by a uniform distributed random variable in the range of $[0.8, 1.2]$. For each run, we collect the optimization run time τ as well the objective J, where J is computed by applying J to the solution obtained by each of the methods.

In Table I we report the mean μ_τ and variance σ_τ. Additionally, we report the average objective function divided by the total demand. i.e., $\bar{J} = J/(\sum_{\mathbf{w}\in W} d_{\mathbf{w}})$. All the scenarios studied were performed using an Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz and 32 GB of RAM memory. The NLP problem was formulated using the IPOPT solver [23], whereas the QP and LP programs were solved using Gurobi 8.1.1 [24].

As expected, we observe that the disjoint model is the slowest, given that its first step requires solving a NLP (followed by a significantly faster solution of an LP). This method takes between 25 and 100 times more time than to



(a) EMA, no rebalancing **(b)** EMA, rebalancing

(c) NYC, no rebalancing **(d)** NYC rebalancing

Fig. 6: Travel times for AMoD users, private vehicles and all vehicles (total) for different penetration rates of AMoDs in the network.

solve CARS3 for EMA and NYC, respectively. Moreover, given that CARS3 requires more variables and constraints, it takes around 30% more time than CARS to solve.

Furthermore, our results of $\bar{J}$ show that the Disjoint method finds the best solution between the three models. The reason for this is that its model for routing is not an approximation. Nevertheless, the solutions of CARS and CARS3 are not more than 4% and 2% away from the Disjoint solution, respectively. Arguably, this result might suggest that the benefit of solving the problem jointly is not as valuable as assumed, which coincides with the results of [9]. However, it is worth mentioning that these results are sensitive to different OD demand distributions. As an example, for perfectly symmetrical OD demands, rebalancing plays no role in the optimization process.

C. System-optimal Routing and Rebalancing Trade-off

Considering the existence of selfish privately-owned vehicles and centrally-controlled AMoD vehicles, we analyze the trade-off that exists between system-optimal AMoD routing and the additional traffic due to AMoD rebalancing in terms of average travel times. We tackle the bi-level Problem (7) following the iterative methodology presented in Section II-D. We use different penetration rates of AMoD customers with respect to the total demand. More specifically, we let $\gamma \in [0, 1]$ be the penetration rate and $\mathbf{g}$ the total OD demand. Then, we assume that $\mathbf{g}^u = \gamma \mathbf{g}$ and $\mathbf{g}^p = (1 - \gamma)\mathbf{g}$ are the AMoD's and private vehicles' demand, respectively. For the purpose of this paper, we choose the same demand distribution for AMoD and private vehicles. Nonetheless, without loss of generality, different demand separation criteria can be readily implemented in this framework.

As shown in Figs. 6a and 6c, the introduction of AMoD users into the system not only improves the overall travel time of AMoD users themselves, but reduces the travel time of private vehicles even more. This is because smart routing decisions of AMoD vehicles reduce the traffic intensity on the congested roads, which consequently allows private vehicles to travel faster. As AMoD users begin to enter the system, we see that the average travel time per vehicle starts

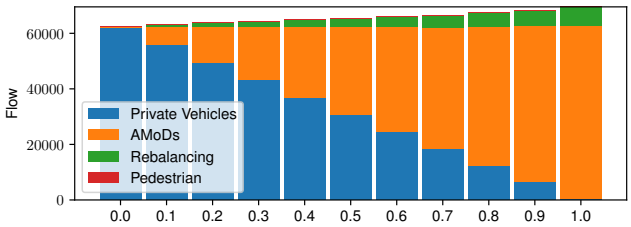


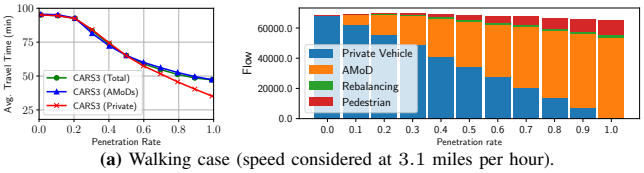
Fig. 7: Distribution of flow in EMA per mode of transport. The x-axis represents the penetration rate of AMoD users.

to decrease compared to the uncontrolled traffic. Moreover, the travel time of commuting through the fastest route (private vehicles) decreases as more AMoD users are in the system.

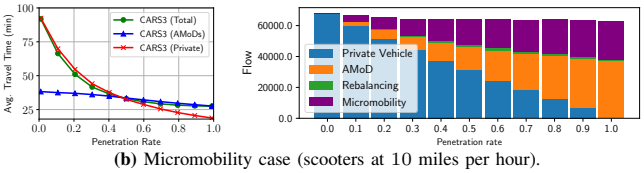
Fig. 6 shows the interaction between the two classes of vehicles when rebalancing is used or not. Comparing Fig. 6b with Fig. 6a, we see that increasing the number AMoD users (penetration rates from 0 to 0.5), all vehicles decrease their travel time. However, as penetration increases (0.5 to 1), a larger amount of vehicles needs to be rebalanced, resulting in a rise of travel times as the overall flow in the network increases as in Fig. 7. The EMA network achieves lower benefits by using system-centric strategies, possibly because the EMA is a highway network with less degrees of freedom in terms of routing decisions than an urban setting. In contrast, for NYC, the impact of rebalancing is negligible, and increasing the number of AMoD users allows to reduce travel time by up to 10%. Notably, these results are in line with the low-to-medium congestion cases in the peak hour presented in [9, Sec. 5.2]. Finally, although the shapes for EMA and NYC shown in Figs. 6b and 6d are not identical, they follow similar trends. In particular, for the 100% AMoD penetration, rebalancing slightly increases the overall travel times for both networks. Yet, in general, the impact of rebalancing on the system-level performance depends on the network topology, and on the symmetry of the OD demand distribution.

D. Walking and Micromobility Options

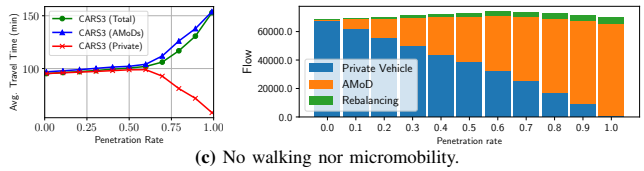
In order to study the impact of centralized routing under high congestion levels, we run experiments for the NYC network with a higher overall demand level (2.5 times higher than in Fig. 6c). As in the previous experiment, we run the analysis for different penetration rates. Notably, the initial travel times shown in Fig. 8 are in line with the high congestion case in the peak hour in [9, Sec. 5.2]. We observe that leveraging the possibility of walking (Fig. 8a), the decrease in the overall travel time is much higher for higher demands, in fact, for a 100% penetration rate it reaches half of travel time compared to a 0% penetration rate. In addition, we consider the option of using micromobility vehicles. In particular, we analyze the case when electric scooters are available to AMoD users everywhere, for which we assume an average speed of 10 miles per hour and the same network as the walking network $\mathcal{G}_{\mathrm{W}}$. Fig. 8b shows that the average travel time for an AMoD user is lower than for selfish users when penetration rates are low. This happens because even for 0% penetration rate, AMoD users resort to scooters which are not available for private vehicles. Similar to other examples, the travel times for both scooters and private



(a) Walking case (speed considered at 3.1 miles per hour).



(b) Micromobility case (scooters at 10 miles per hour).



(c) No walking nor micromobility.

Fig. 8: Effect of alternative mode of transport in NYC when demand is high. We increase demand by a 2.5 factor, i.e., we use $2.5\mathbf{g}$.

vehicles decrease as the penetration rate increases. Finally, we explore the same setting as before but without walking or micromobility options, with the purpose of observing the effect of including these services on an AMoD system. In Fig. 8c, we observe that the injection of AMoD users to the network increases travel times. This is a result of the additional rebalancing flow needed to operate the system in high demand periods, and happens due to the evaluation of $t(\cdot)$ at those points: Every additional flow increases travel times quartically when congestion is high.

In conclusion, by comparing Fig. 8c with Fig. 8a and 8b, we see that pure AMoD systems might decrease the system-level performance due to the additional congestion resulting from rebalancing the AMoD vehicles. Yet, combining centralized-routing with the possibility of walking or using micromobility solutions such as e-scooters can significantly improve the overall travel times.

V. Conclusions

In this paper we studied the achievable benefits of centrally controlling an Autonomous Mobility-on-Demand (AMoD) system under mixed traffic conditions. With the goal of minimizing the customers' travel time, we extended a previously presented quadratic model [10] by improving its accuracy and included a reactive exogenous traffic flow. Assuming the exogenous traffic (private vehicles) to act selfishly, we leveraged an iterative method [11] to study the interaction between AMoD and private cars. Finally, we presented numerical experiments to compare the proposed method with the state-of-the-art, and to gain insights on the achievable benefits for different AMoD penetration rates and micromobility options. Our results showed that the proposed method outperforms the state-of-the-art, and revealed that combining AMoD rides with walking and micromobility options can significantly improve the overall system-level performance.

This work can be extended as follows. First, given the large computational time of the disjoint problem (NLP) we would like to propose a MSA-type method to solve the fleet-interested social-centric TAP considering exogenous flow, possibly leveraging computationally efficient algorithms such as in [25]. Second, we would like to generalize the approximation model to n line segments, and provide theoretical bounds on the model error. Third, given that the solution of these models are in terms of flow, we would like to include route-recovery strategies and apply this framework to larger networks, and to test our method with high-fidelity simulations. Finally, we would like to consider a more general intermodal setting as in [26], [27] by including public transportation options.

References

[1] J. G. Wardrop, "Road paper. some theoretical aspects of road traffic research." *Proceedings of the institution of civil engineers*, vol. 1, no. 3, pp. 325–362, 1952.

[2] J. Guanetti, Y. Kim, and F. Borrelli, "Control of connected and automated vehicles: State of the art and future challenges," *Annual Reviews in Control*, vol. 45, pp. 18–40, 2018.

[3] R. M. Swaszek and C. G. Cassandras, "Load balancing in mobility-on-demand systems: Reallocation via parametric control using concurrent estimation," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 2148–2153.

[4] S. Hörl, C. Ruch, F. Becker, E. Frazzoli, and K. W. Axhausen, "Fleet control algorithms for automated mobility: A simulation assessment for Zurich," in *Annual Meeting of the Transportation Research Board*, 2018.

[5] M. W. Levin, 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, pp. 373 – 383, 2017.

[6] R. Zhang and M. Pavone, "Control of robotic Mobility-on-Demand systems: A queueing-theoretical perspective," *Int. Journal of Robotics Research*, vol. 35, no. 1–3, pp. 186–203, 2016.

[7] R. Iglesias, F. Rossi, R. Zhang, and M. Pavone, "A BCMP network approach to modeling and controlling Autonomous Mobility-on-Demand systems," in *Workshop on Algorithmic Foundations of Robotics*, 2016.

[8] M. Pavone, 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, pp. 839–854, 2012.

[9] F. Rossi, R. Zhang, Y. Hindy, and M. Pavone, "Routing autonomous vehicles in congested transportation networks: Structural properties and coordination algorithms," *Autonomous Robots*, vol. 42, no. 7, pp. 1427–1442, 2018.

[10] M. Salazar, M. Tsao, I. Aguiar, M. Schiffer, and M. Pavone, "A congestion-aware routing scheme for autonomous mobility-on-demand systems," in *European Control Conference*, 2019, in press.

[11] A. Houshmand, S. Wollenstein-Betech, and C. G. Cassandras, "The penetration rate effect of connected and automated vehicles in mixed traffic routing," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1755–1760.

[12] M. J. Beckmann, C. B. McGuire, and C. B. Winsten, "Studies in the Economics of Transportation," p. 359, 1955.

[13] S. Wollenstein-Betech, C. Sun, J. Zhang, and I. C. Paschalidis, "Joint estimation of od demands and cost functions in transportation networks from data," *arXiv preprint arXiv:1909.00941*, 2019.

[14] Bureau of Public Roads, "Traffic assignment manual," U.S. Dept. of Commerce, Urban Planning Division, Tech. Rep., 1964.

[15] M. Patriksson, *The traffic assignment problem: models and methods*. Courier Dover Publications, 2015.

[16] C. F. Daganzo and Y. Sheffi, "On stochastic models of traffic assignment," *Transportation science*, vol. 11, no. 3, pp. 253–274, 1977.

[17] Y. A. Korilis, A. A. Lazar, and A. Orda, "Achieving network optima using stackelberg routing strategies," *IEEE/ACM transactions on networking*, vol. 5, no. 1, pp. 161–173, 1997.

[18] D. A. Lazar, S. Coogan, and R. Pedarsani, "Capacity modeling and routing for traffic networks with mixed autonomy," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 5678–5683.

[19] F. Rossi, "On the interaction between Autonomous Mobility-on-Demand systems and the built environment: Models and large scale coordination algorithms," Ph.D. dissertation, Stanford University, Dept. of Aeronautics and Astronautics, 2018.

[20] P. T. Harker, "Multiple equilibrium behaviors on networks," *Transportation science*, vol. 22, no. 1, pp. 39–46, 1988.

[21] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org , 2017.

[22] Data retrieved from Uber Movement, "2019 Uber Technologies, Inc," 2019. [Online]. Available: https://movement.uber.com
[23] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical programming*, vol. 106, no. 1, pp. 25–57, 2006.
[24] L. Gurobi Optimization, "Gurobi optimizer reference manual," 2020. [Online]. Available: http://www.gurobi.com
[25] K. Solovey, M. Salazar, and M. Pavone, "Scalable and congestion-aware routing for autonomous mobility-on-demand via frank-wolfe optimization," in *Robotics: Science and Systems*, 2019.
[26] M. Salazar, 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, Extended Version, Available at https://arxiv.org/abs/1804.11278.
[27] M. Salazar, N. Lanzetti, F. Rossi, M. Schiffer, and M. Pavone, "Intermodal autonomous mobility-on-demand," *IEEE Transactions on Intelligent Transportation Systems*, 2019.