A Model Predictive Control Scheme for Intermodal Autonomous Mobility-on-Demand

Jannik Zgraggen[1,2], Matthew Tsao[2], Mauro Salazar[2], Maximilian Schiffer[2,3] and Marco Pavone[2]

***Abstract*— This paper presents a routing algorithm for intermodal Autonomous Mobility on Demand (AMoD) systems, whereby a fleet of self-driving cars provides on-demand mobility in coordination with public transit. Specifically, we present a time-variant flow-based optimization approach that captures the operation of an AMoD system in coordination with public transit. We then leverage this model to devise a model predictive control (MPC) algorithm to route customers and vehicles through the network with the objective of minimizing customers' travel time. To validate our MPC scheme, we present a real-world case study for New York City. Our results show that servicing transportation demands jointly with public transit can significantly improve the service quality of AMoD systems. Additionally, we highlight the differences of our time-variant framework compared to existing mesoscopic, time-invariant models.**

I. Introduction

Road congestion causes billions of dollars of annual economic losses resulting from the time people spend stuck in traffic and health issues due to pollution. Currently, this loss ranges in between 83 to 110 billion dollars annually, excluding additional externalities such as threats to public health and environmental harm caused by pollution [1]. Experts envision the severity of these losses to increase in the future due to population growth and increased urbanization [2].

Resolving the congestion problem without disrupting everyday mobility services preoccupies municipalities as well as mobility providers. Experts agree that a paradigm shift in mobility services is necessary to address congestion, but compatibility with the existing infrastructure heavily constrains potential solutions. Additionally, current trends, such as ride-hailing services, are disrupting the mobility landscape by offering low-cost urban road transport, which increases congestion even further due to induced demands. This cheaper, more comfortable service shifts customer demands from public transit to road transport. In Manhattan, the number of for-hire vehicles raised from 47,000 to 103,000 between 2003 and 2018, while the average traffic speed decreased from 6.5 mph to 4.7 mph [3] and public transit usage dropped for the first time in history [4].

The advent of self-driving technology opens the door for many possibilities towards new solutions that can help to mitigate congestion problems. In current mobility systems, each agent (e.g., a single taxi driver, or a company like Uber), aims to maximize its own profit without cooperating with other service providers. This lack of cooperation between service providers leads to global inefficiencies, resulting in increased traffic and thus additional congestion. In contrast, systems that merge the concepts of self-driving vehicles and ride-hailing also known as Autonomous Mobility-on-Demand (AMoD) systems, can centrally control the autonomous service vehicles in line with a global objective, e.g. in a congestion-aware fashion.

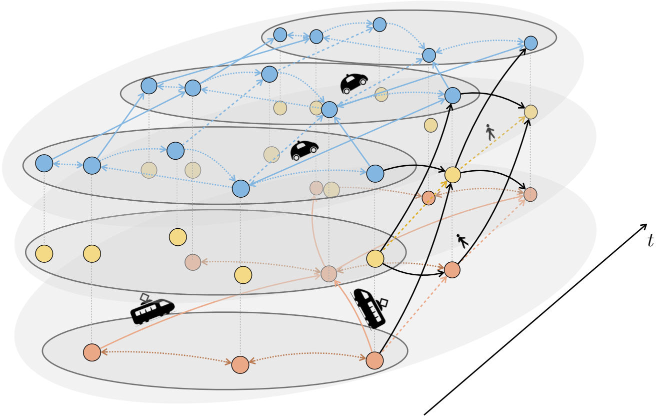

Fig. 1. The time-varying intermodal AMoD network consists of a time-expansion of a road digraph (blue), a public transit digraph (orange) and pick-up/drop-off nodes (yellow). The colored dots denote intersections or stops and the arrows represent sample connections. Specifically, dotted arrows denote geographical links and grey dotted lines denote geographically close nodes. Colored solid arrows represent road and public transit connections, whilst the colored dashed lines denote waiting links. Black solid arrows represent mode switching arcs.

In such a system, a central operator assigns passenger requests to the vehicles and coordinates re-balancing routes to align the position of empty vehicles with upcoming transportation demand [5]. AMoD systems feature several advantages, which allow for mobility services at lower prices and increased availability. However, an isolated AMoD system may itself cause induced demand and cannibalize other means of transport. Thus while isolated AMoD systems may be more efficient than existing solutions, they may not be sufficient to resolve congestion problems in a sustainable fashion. Efficient means of public transport that are less convenient, such as trains and buses, may experience a strong reduction in utilization due to this induced demand effect.

Accordingly, efficient and sustainable large-scale deployment of AMoD will only be possible if vehicle fleets interact intelligently with existing public transit infrastructure, to support a sustainable utilization of both systems and to avoid demand cannibalization. Preliminary mesoscopic studies showed that an intelligent interaction between AMoD and public transit in the form of intermodal AMoD (I-AMoD) can yield significant benefits compared to an AMoD system

[1] Automatic Control Laboratory, EPFL, Lausanne, Switzerland `jannik.zgraggen@epfl.ch`

[2] Autonomous Systems Lab, Stanford University, Stanford (CA), United States {`mwtsao,samauro,pavone`}`@stanford.edu`

[3] TUM School of Management, Technical University of Munich, Munich, Germany `schiffer@tum.de`

operating in isolation [6], [7]. In this paper, we develop a routing algorithm which allows to control such a system in practice and futher enables us to verify and refine the findings of previous time-invariant studies.

Related literature: Our work contributes to two different research streams, namely, control of AMoD systems, and intermodal passenger transport. In the following, we review these research streams.

Several approaches exist to model and control AMoD systems, ranging from queuing-theoretical models [8] to simulation-based models [9], [10], [11] and multi-commodity network flow models [12], [13]. Queueing-theoretical models capture the stochasticity of the customer arrival process and are amenable to efficient control synthesis. However, their complex structure is not well suited to capture the interaction with other modes of transportation. These models also assume that the demand distribution is time invariant, which may not be accurate in practice. Simulation-based models capture transportation systems with very high fidelity but are generally not amenable to optimization. Network flow models are amenable to optimization and can capture a variety of complex constraints. In fact, they have been widely used in problems ranging from (congestion-aware) route planning schemes for AMoD systems [14], [15], [16], to the joint control of AMoD systems and the electric power network [17], and stochastic model predictive control (MPC) algorithms for single-customer [18], [19] and ride-sharing AMoD [20].

Research on intermodal passenger transportation is still sparse. Existing work on the interplay between AMoD and public transportation is either based on simulation [21], [22], [10] or on fluidic [23] models. However, these studies focus on the analysis of specific scenarios, and do not consider the *optimization* of joint control policies for AMoD systems and public transit. So far, only our previous studies [6], [7] offered a mesoscopic optimization framework for an I-AMoD system but is limited to system analysis in a time-invariant setting.

In summary, some frameworks for the operation of AMoD systems are available but lack the consideration of public transit. Vice versa, frameworks that consider public transit lack an optimization-based routing component.

Statement of contributions: This paper presents a routing algorithm that provides customer and vehicle routes for I-AMoD systems. Specifically, our contribution is threefold: First, we develop a time-variant multi-commodity network flow optimization model that captures the joint operation of AMoD systems and public transit (cf. Fig. 1). To increase social welfare, our objective comprises a combination of customers' travel time and operational costs. Additionally, we consider congestion effects by capturing the impact of exogenous traffic on travel time and accordingly limiting additional transit delays induced by the operation of an AMoD fleet. Second, we propose a high-level MPC scheme which periodically solves a time-variant network flow optimization problem and samples routes from the solution in a receding-horizon fashion to incorporate new information as it is revealed. Third, we present a real-world case study for Manhattan which we use to test the proposed controller and compare its performance to an AMoD system operating in isolation. We show that the the total time spent in traffic can be reduced by up to 25 % by jointly coordinating public transit and AMoD. Additionally, we analyze differences between our time-variant results and time-invariant results from previous studies.

Organization: The remainder of this paper is structured as follows: In Section II we present a multi-commodity network flow optimization model for time-variant I-AMoD, which we leverage via MPC to produce a routing algorithm for I-AMoD in Section III. In Section IV, we introduce our case study for Manhattan and present simulation results. Finally, Section V concludes this paper with a short summary and an outlook on future research.

II. Model

This section introduces a flow optimization model for time-variant I-AMoD systems. We present a centralized system with the following functionalities: i) it assigns transportation requests to services, ii) it considers multiple modes of transportation, namely walking, subway, and AMoD vehicles, iii) it respects road capacity limits, whilst accounting for exogenous and endogenous congestion, and iv) it rebalances empty vehicles to re-align their distribution with future transportation demand. Section II-A introduces a multi-commodity network flow optimization model. We describe a model for congestion in Section II-B, whilst we present a time and space clustering approach in Section II-C.

A. Time-variant Intermodal Network Flow Model

We represent the intermodal transportation network as a directed graph $\mathscr{G} = (\mathscr{V}, \mathscr{A})$, with a set of arcs $\mathscr{A}$ and a set of vertices $\mathscr{V}$. Time is to a resolution of τ for a finite number of n time steps such that the time horizon is

$$\mathscr{T}(t_0,n) := \{t_0+\tau, t_0+2\tau, ..., t_0+n\tau\}, \tag{1}$$

with the current time t_0 and a length of $n\tau$. We use a time-expanded graph representation, i.e., every vertex $j = (t_j, l_j) \in \mathscr{V}$ is characterized by its geographical location l_j and a time index $t_j \in \mathscr{T}(t_0,n)$. An arc exists between two vertices $i = (l_i,t_i)$ and $j = (l_j,t_j)$ if a transportation mode (i.e., walking, subway, or AMoD) can depart from location l_i at time t_i and arrive at l_j at time t_j. Due to the time discretization, the travel time of each arc is a multiple of τ.

The transportation network $\mathscr{G}$ has three modes of transportation: walking, AMoD vehicles and subway. Accordingly, we partition vertices into three sets such that $\mathscr{V} = \mathscr{V}_\mathrm{W} \cup \mathscr{V}_\mathrm{R} \cup \mathscr{V}_\mathrm{P}$, with $\mathscr{V}_\mathrm{W}, \mathscr{V}_\mathrm{R}, \mathscr{V}_\mathrm{P}$ representing the walking, road, and public transit nodes of the network, respectively. Arcs within $\mathscr{V}_\mathrm{W}, \mathscr{V}_\mathrm{R}, \mathscr{V}_\mathrm{P}$ represent movement via the corresponding mode, and arcs between the sets $\mathscr{V}_\mathrm{W}, \mathscr{V}_\mathrm{R}, \mathscr{V}_\mathrm{P}$ represent changing modes of transportation. Accordingly, we partition arcs $\mathscr{A} = \mathscr{A}_\mathrm{cus} \cup \mathscr{A}_\mathrm{reb} \cup \mathscr{A}_\mathrm{veh}$ into a set $\mathscr{A}_\mathrm{cus}$ comprising all arcs that denote sidewalks and subway lines on which customers can walk, ride the public transit or just wait; a set $\mathscr{A}_\mathrm{reb}$ used to signify when an empty vehicle is en route to pick up a customer or has just dropped off a customer; and a set $\mathscr{A}_\mathrm{veh}$ that denotes vehicle flow arcs between different regions. To fully specify the intermodal structure in such a graph, we further categorize arcs into intra-regional and inter-regional arcs. These arcs describe mode-switching, waiting, pick-up and drop-off (intra-regional, see Fig. 2), and transportation (inter-regional). To keep this paper concise, we exhaustively explain this concept in Appendix A.

To formulate the I-AMoD system as a multi-commodity network, we represent consumers and vehicles as commodities. To this end, we introduce two different types of commodities: service vehicles and transportation demand. Each commodity is defined by a sink, a source, and a quantity. The transportation demand of the system is given by the set of all travel request commodities $\mathscr{R}$. There are M different request commodities. Each transportation request $r_m = (o_m, d_m, t_m, a_m) \in \mathscr{R}$ for some $m \in \{1, 2, ..., M\} =: \mathscr{M}$ is a 4-tuple specifying its geographical location of the origin o_m and the destination d_m, the time of the request t_m, and the number of customers a_m associated with it. Note that the source of a request commodity is given by a vertex $i = (o_m, t_m)$, whilst its sink is given by all vertices $j = (d_m, t) \quad \forall t \geq t_m \in \mathscr{T}(t_0, n)$. We denote commodity flow variables as $f_m(i,j)$ and $f_0(i,j)$, where $f_m(i,j)$ represents the number of customers of request type m traveling on arc $(i,j) \in \mathscr{A}$ and $f_0(i,j)$ denotes the number of empty vehicles moving on arcs $(i,j) \in \mathscr{A}$ to rebalance vehicles. We define $\{a_m^t\}_{t \in T}$ so that a_m^t is the number of type m requests that are delivered at time t. With this notation, we state the I-AMoD routing objective:

$$
J\Big(\{f_m(\cdot,\cdot)\}_m, f_0(\cdot,\cdot)\Big) := \tag{2}
$$
$$
\Big(\sum_{(i,j)\in\mathscr{A}} \sum_{m=1}^{M} \rho_{ij}^t \cdot f_m(i,j)\Big) + \Big(\sum_{(i,j)\in\mathscr{A}} \rho_{ij}^o \cdot \Big[f_0(i,j) + \sum_{m=1}^{M} f_m(i,j)\Big]\Big).
$$

The objective function in (2) penalizes customer inconvenience and operation cost. The constants ρ_{ij}^t represent the customer costs of traversing arc (i,j). Similarly, the constant ρ_{ij}^o denotes the cost of moving vehicles on across arc (i,j).

The routing strategy $\{f_m(\cdot,\cdot)\}_m, f_0(\cdot,\cdot)$ must satisfy the following constraints: i) flow non-negativity, ii) road capacity constraints, iii) conservation of customers, iv) request completion and v) conservation of vehicles. Formally, these constraints hold as follows:

$$
f_m(i,j) \geq 0 \qquad \forall\, m \in \{0\} \cup \mathscr{M} \tag{3}
$$

$$
f_0(i,j) + \sum_{m\in\mathscr{M}} f_m(i,j) \leq c_{ij} \qquad \forall (i,j) \in \mathscr{A}_{\text{veh}} \cup \mathscr{A}_{\text{reb}} \tag{4}
$$

$$
\sum_{\substack{i:(i,j)\\ \in\mathscr{A}_{\text{cus}}\setminus\mathscr{A}_{\text{reb}}}} f_m(i,j) + \mathbb{1}_{\{o_m=l_j, t_m=t_j\}} a_m = \sum_{\substack{k:(j,k)\\ \in\mathscr{A}\setminus\mathscr{A}_{\text{reb}}}} f_m(j,k) + \mathbb{1}_{d_m=l_j} a_m^{t_j}
$$
$$
\forall m \in \mathscr{M},\ j \in \mathscr{V} \tag{5}
$$

$$
a_m = \sum_{t=1}^{|\mathscr{T}(t_0,n)|} a_m^t \qquad \forall m \in \mathscr{M} \tag{6}
$$

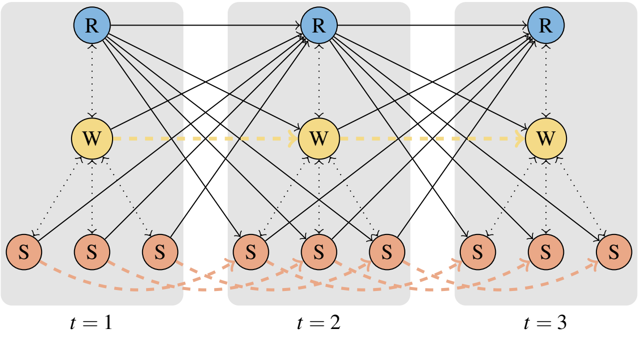

Fig. 2. Illustration of intra-regional arcs at three different time steps. Inside the region there are three public transportation stations. Colored dashed arrows represent waiting arcs. Solid black arrows denote AMoD pickup or delivery, whilst dotted black arrows are mode-switching arcs for rebalancing vehicles.

$$
\sum_{m\in\mathscr{M}} \sum_{i:(i,j)\in\mathscr{A}_{\text{veh}}} f_m(i,j) + \sum_{\substack{i:(i,j)\\ \in\mathscr{A}_{\text{reb}}\cup\mathscr{A}_{\text{veh}}}} f_0(i,j) + \mathscr{D}_{\text{initial}} = \tag{7}
$$
$$
\sum_{m\in\mathscr{M}} \sum_{k:(j,k)\in\mathscr{A}_{\text{veh}}} f_m(j,k) + \sum_{\substack{k:(j,k)\\ \in\mathscr{A}_{\text{reb}}\cup\mathscr{A}_{\text{veh}}}} f_0(j,k) + \mathscr{D}_{\text{final}} \quad \forall j \in \mathscr{V}.
$$

Non-negative flows: Constraint (3) enforces flow variables to be positive, as commodities can only move forward in time.

Road capacity constraints: The inequality constraint (4) ensures that the total number of vehicles traversing any arc $(i,j) \in \mathscr{A}$ cannot exceed its capacity c_{ij}.

Conservation of customers: Customers cannot appear or disappear at locations other than their origins and destinations, respectively. For each trip type m, constraint (5) ensures that each customer entering a node must leave it.

Request completion: The system should serve all customers who request rides. Since a_m is the total number of type m requests, the constraint to serve all customers is then captured by constraint (6).

Conservation of vehicles: Since the number of AMoD vehicles is fixed, the constraint (7) ensures that each vehicle entering a node $j \in \mathscr{V}$ has to exit it. Recalling the definition of $\mathscr{A}_{\text{veh}}$, a customer is traveling on an arc in $\mathscr{A}_{\text{veh}}$ if and only if they are riding in an AMoD vehicle. Therefore the flow f_m restricted to $\mathscr{A}_{\text{veh}}$ is precisely the flow of customer carrying AMoD vehicles. The terms $\mathscr{D}_{\text{initial}}, \mathscr{D}_{\text{final}}$ specify the initial and final distributions of the vehicles with respect to the planning horizon.

With the objective function and system constraints formalized, we now present the optimization problem for controlling an I-AMoD system:

$$
\underset{\{f_m(\cdot,\cdot)\}_m, f_0(\cdot,\cdot)}{\text{minimize}} \; J\Big(\{f_m(\cdot,\cdot)\}_m, f_0(\cdot,\cdot)\Big) \tag{8}
$$
$$
\text{s.t. } (3), (4), (5), (6), (7).
$$

Problem (8) is a linear program (LP) and can be solved efficiently using interior point methods.

B. Congestion Model

Congestion influences travel times on road arcs. We assume the I-AMoD fleet to be significantly smaller than the number of privately owned vehicles on the road so that the road congestion levels are approximately independent of the actions of the I-AMoD fleet. This is to say that the road congestion is well approximated by the *exogenous* congestion, which is independent of the I-AMoD fleet. We use the Bureau of Public Roads (BPR) function [24] of the form $F_{\mathrm{BPR}}(x) = 1 + 0.15x^4$ to calculate congestion dependent travel times. Here x represents the ratio between the vehicle flow traversing a road link and its nominal capacity. With this model the time it takes to traverse the link (l_i, l_j) can be written as

$$t_{l_il_j} = t^{\mathrm{N}}_{l_il_j} F_{\mathrm{BPR}}\left(u^{\mathrm{R}}_{l_il_j} / c^{\mathrm{R}}_{l_il_j}\right), \tag{9}$$

where $t^{\mathrm{N}}_{l_i,l_j}$ is the free flow traversal time on link (l_i, l_j), $u^{\mathrm{R}}_{l_i,l_j}$ the exogenous traffic flow and $c^{\mathrm{R}}_{l_i,l_j}$ its nominal capacity. The endogenous congestion effect can either be modeled with a penalty term in the objective function or by the presence of capacity constraints. As proposed in [7], we choose to set the capacity limits of our system in such a way that the AMoD traffic does not increase travel time more than a factor Δr_{time}. Formally, it holds that

$$t_{l_il_j} = t^{\mathrm{N}}_{l_il_j} \cdot F_{\mathrm{BPR}}\left(c^{\mathrm{R,th}}_{l_il_j} / c^{\mathrm{R}}_{l_il_j}\right). \tag{10}$$

Therefore, the capacity available to the I-AMoD system is

$$c_{l_il_j} = \left(\frac{\Delta r_{\mathrm{time}}}{0.15} + \left(\frac{u^{\mathrm{R}}_{l_il_j}}{c^{\mathrm{R}}_{l_il_j}}\right)^4\right)^{\frac{1}{4}} c^{\mathrm{R}}_{l_il_j} - u^{\mathrm{R}}_{l_il_j} \;\; \forall (i,j) \in \mathscr{A}_{\mathrm{veh}} \cup \mathscr{A}_{\mathrm{reb}}. \tag{11}$$

This way, we can calculate the travel times and road capacities once the values of $c^{\mathrm{R}}_{l_il_j}, u^{\mathrm{R}}_{l_il_j}, \Delta r_{\mathrm{time}}$ are specified.

C. Clustering the road network

We partition the road network into regions for two main reasons: First, travel times need to be multiples of τ to be accurately represented in $\mathscr{G}$; second, such a partitioning additionally limits the size of $\mathscr{G}$ and keeps the I-AMoD optimization problem computationally tractable.

We propose a `Greedy Clustering Heuristic` to partition the nodes of a high resolution network $\mathscr{G}_0$ whose nodes are locations (such as street intersections or public transit stops) and arcs are roads or public transit routes as follows: First, we fix a cluster radius r and initialize a set of centroids to the empty set $\mathscr{N} \leftarrow \emptyset$. Then, we check if there exists a vertex in $\mathscr{G}_0$ with distance greater than r to all vertices in $\mathscr{N}$. If such a vertex exists, we add it to $\mathscr{N}$ as a centroid; otherwise, our clustering terminates and we assign each vertex in $\mathscr{G}_0$ to the cluster representing its closest centroid. We calculate the distance between clusters as the mean distance between their vertices in $\mathscr{G}_0$ and the capacity between the clusters by the sum of capacities of the direct links between two clusters in $\mathscr{G}_0$.

Recall that time is discretized into time units of size τ, and that this can cause rounding errors for travel times if they are not multiples of τ. Specifically, travel times are rounded up to the next multiple of τ and thus the rounding error for a travel time t is given by $\varepsilon_\tau(t) := \tau\lceil \frac{t}{\tau} \rceil - t$. To keep the rounding errors small, for a specified congestion level we run `Greedy Clustering Heuristic` with a radius r_τ where r_τ is the distance a vehicle can travel in τ time for clustering.

III. Model Predictive Control Scheme

The LP as presented in Section II is not directly applicable as a routing strategy, for three main reasons. First, the model assumes perfect information about future demand, which is not true in practice. Second, the problem becomes intractable if the optimization horizon is too large. Third, the solution to problem (8) is fractional, but an integer solution is required to operate a transportation system. We address these issues in Sections III-A, III-B and III-C, respectively. Based on this, in Section III-D we present an MPC scheme for the operation of I-AMoD systems based on Problem (8) in Section II-A.

A. Forecasting Customer Demand

In the absence of perfect information on future customer demand, estimates of the demand can be used instead. Machine learning models are used to forecast travel demand for the near future based on the recent history of the system in [18],[19],[20]. They show that accurate forecasting models can be learned on a wide range of taxi and transportation data sets, and that leveraging these forecasts can give significant improvements to customer waiting times compared to reactive algorithms, i.e. algorithms that do not act on estimates of future information. Following this methodology, for a time horizon $\mathscr{T}$, let $\Lambda_{\mathscr{T}}$ denote the demand that appears in the horizon. Similarly, we define $\widehat{\Lambda}_{\mathscr{T}}$ to be a demand *forecast* for $\mathscr{T}$ that serves as an estimate of $\Lambda_{\mathscr{T}}$. Demand that appears at the very end of the horizon may be impossible to deliver within the horizon. For this reason, we use two time horizons $\mathscr{T}_{\mathrm{pred}}(t) := \mathscr{T}(t,m)$ and $\mathscr{T}_{\mathrm{opt}}(t) := \mathscr{T}(t,n)$ where t is the current time and $m < n$ are integers specifying the lengths of the horizons as per (1). To assimilate the forecasting framework into the model proposed in Section II-A, we solve (8) with a time horizon $\mathscr{T}_{\mathrm{opt}}(t)$ where the requests are given by $\widehat{\Lambda}_{\mathscr{T}_{\mathrm{pred}}(t)}$. Since $\mathscr{T}_{\mathrm{opt}}(t)$ is longer than $\mathscr{T}_{\mathrm{pred}}(t)$, the optimizer has time to schedule requests appearing near the end of the prediction horizon $\mathscr{T}_{\mathrm{pred}}(t)$.

B. Computational Tractability

We address computational tractability by reducing the lengths of $\mathscr{T}_{\mathrm{pred}}(t)$ and $\mathscr{T}_{\mathrm{opt}}(t)$. Using a shorter time horizon, however, reduces the amount of information that can be used in the optimization. To address this, we periodically update the forecast $\widehat{\Lambda}_{\mathscr{T}_{\mathrm{pred}}(t)}$ to incorporate new information and resolve (8) in a receding-horizon fashion. In such an approach, (8) may become infeasible if $\widehat{\Lambda}_{\mathscr{T}_{\mathrm{pred}}(t)}$ contains too many customers who cannot all be served within $\mathscr{T}_{\mathrm{opt}}(t)$. Accordingly, constraint (6) might be violated. To preserve feasibility, we allow customers to be dropped and remove constraint (6). To avoid the trivial solution with all customers dropped, we reformulate (2) by adding a drop penalty term to incentivize serving as many customers as possible. Accordingly, the softconstrained I-AMoD optimization problem is

$$\begin{aligned}\underset{\{f_m(\cdot,\cdot)\}_m, f_0(\cdot,\cdot)}{\text{minimize}} \quad & J\Big(\{f_m(\cdot,\cdot)\}_m, f_0(\cdot,\cdot)\Big) \\ & + \sum_{m=1}^{M}\Big(\sum_{\substack{(i,j)\in\mathscr{A}\\ \text{s.t. } l_j\neq m\\ t_j=|\mathscr{T}_{\text{opt}}(t)|}} P\cdot f_m(i,j)\Big) \\ \text{s.t. } & (3),(4),(5),(7).\end{aligned} \tag{12}$$

The second term in the objective imposes a drop penalty P for any request that has not reached its destination by the end of the planning horizon.

To further reduce the computational complexity of (12) we bundle customer flows and distinguish them by their geographical destination [27]. To translate flows back to the 4-tuple request commodities, we apply the post-processing flow decomposition algorithm presented in [27] to decompose the optimal solution $(f_0^\star, f_m^\star)$ into a set of K routes. A route is given by a tuple $(r_k^m, \alpha_k^m) \in \mathscr{R}^\star \quad \forall k = 1,2,\ldots,K$, where r_k^m is the sequence of nodes on route k, α_k^m is the customer flow on route k, and m is the corresponding unbundled customer commodity. This significantly reduces the number of commodity flow variables, whilst not changing the optimal solution of the problem. In the flow bundling formulation, the problem size does not depend on the number of vehicles or customers making it suitable for a large scale settings.

C. Fractional Flows

The flow values α_k^m obtained from the solution of the LP are noninteger. To obtain an integer routing from $\mathscr{R}^\star$ whilst incurring low rounding errors, we use flow decomposition to extract as many integer flows from $\mathscr{R}^\star$ as possible. After one route k is chosen, we update the residual flow (r_k^m, α_k^m) with $(r_k^m, \alpha_k^m - 1)$ in the set $\mathscr{R}^\star$. If for a given demand m there are no remaining routes with $\alpha_m^k \geq 1$ we sample one route r_k^m randomly from the distribution proportional to the remaining flow. In practice the solution of (8) and the rounded version described here are very similar, but in principle it is not clear how different the objective values of the rounded and original solutions can be in general.

D. Algorithm

We denote $S^t(\mathscr{D}^t, \mathscr{C}^t, \widehat{\Lambda}_{\mathscr{T}_{\text{opt}}(t)})$ as the state of our system at time t. Herein, $\mathscr{D}^t$ denotes the expected spatial and temporal distribution of cars at time t. To account for capacity limits, $\mathscr{C}^t$ denotes the set of expected capacities for each road arc.

We present an MPC algorithm based on (8) which is outlined in Algorithm 1. First, we estimate the system's state, i.e., we generate a forecast $\widehat{\Lambda}_{\mathscr{T}_{\text{pred}}(t)}$, check road congestion levels to obtain $\mathscr{C}^t$ and measure the vehicles' locations to obtain $\mathscr{D}^t$. Second, we solve (8) using the current state as an input to obtain a fractional bundled solution $(f_0^\star, f_m^\star)$. Third, we debundle the solution and convert it into an integer flow by the procedures described in Sections III-B and III-C respectively. Fourth, we execute the integer routing for τ units of time. To address the fact that $\mathscr{T}_{\text{opt}}(t)$ is only a subset of the entire operation horizon, these four steps are repeated every τ units of time so that $\mathscr{D}^t, \mathscr{C}^t, \widehat{\Lambda}_{\mathscr{T}_{\text{opt}}(t)}$ can be periodically updated to incorporate new information observed in the system during operation.

Algorithm 1 Model Predictive Control

```
procedure I-AMoD-MPC
    while $\widehat{\Lambda}_{\mathscr{T}_{\text{pred}}(t)} \neq \emptyset$ do ▷ There is predicted demand
        $S^t \leftarrow$ estimate $\widehat{\Lambda}_{\mathscr{T}_{\text{pred}}(t)}$, $\mathscr{D}^t$ and $\mathscr{C}^t$
        $(f_0^\star, f_m^\star) \leftarrow$ solve LP (12)
        $\mathscr{R}^\star \leftarrow$ flow decomposition of $(f_0^\star, f_m^\star)$
        Route customers and vehicles
```

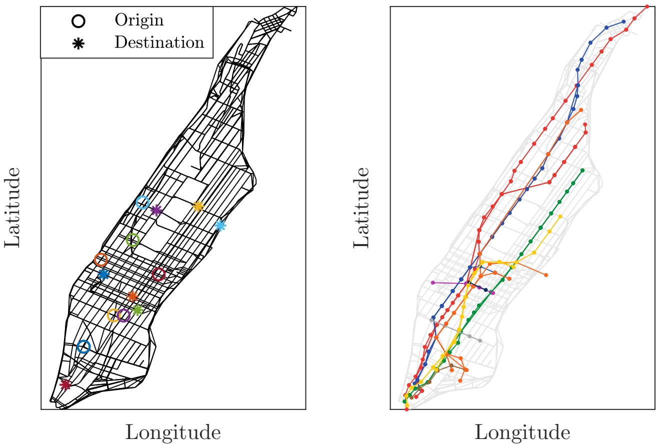

Fig. 3. Road network and subway lines of Manhattan with exemplary origin-destination pairs. Taken from [6].

IV. Numerical Experiments

We validate the proposed MPC Algorithm 1 in a case study of New York City, whereby we simulate an I-AMoD system servicing transportation requests from the taxi & limousine commission data set. Specifically, we introduce the New York City Taxi & Limousine data set and the simulation environment in Section IV-A. Section IV-B discusses the experimental design. We present numerical results of two experiments in Sections IV-C and IV-D, respectively. Section IV-E concludes with a discussion of the results.

A. Simulation Environment

The New York Taxi and Limousine Commission data set contains 53,932 taxi rides served on March 1, 2012 between 6 and 8 pm within Manhattan. Among these trips are 6,772 unique origin-destination pairs. The road network topology is obtained by OpenStreetMap data [28] and the road capacities are proportional to the number of lanes multiplied by the road's speed limit.

We consider the transportation network of Manhattan shown in Fig. 3, consisting of a road network and subway lines. Since the subway is the dominant public transit mode in Manhattan, we use it to build the public transportation network for our case study. We construct the public transportation digraph based on the geographical location of the lines and the stops found in the NYC Open Data database [29]. Our time discretization is $\tau = 2\,\text{min}$ and the time expanded graph is constructed via the `Greedy Clustering Heuristic` described in Section II-C. The choice of clustering radius r_τ from Section II-C ensures that the geographical resolution of the model is proportional to

TABLE I

NUMERICAL DATA FOR THE CASE STUDY

Parameter	Variable	Value	Source
Value of time	V_{T}	24.40 USD/h	[30]
Vehicle operational cost	$V_{\mathrm{D,R}}$	0.486 USD/mile	[32]
Subway operational cost	$V_{\mathrm{D,P}}$	0.47 USD/mile	[31]

the congestion level, allowing for high resolution routing stategies in high congestion situations. In accordance with public transit schedules, the subway can be boarded from a subway station once every 6 min. For the sake of simplicity, we assume that perfect forecasts for the horizon $\mathcal{T}_{\text{pred}}(t)$ are available, i.e., $\widehat{\Lambda}_{\mathcal{T}_{\text{pred}}(t)} = \Lambda_{\mathcal{T}_{\text{pred}}(t)}$. This assumption is without loss of generality as [18], [19], [20] show that *estimates* of future demand also lead to substantial improvement over reactive algorithms in AMoD problems. Finally, we assume exogenous traffic to be known a priori. Therefore, as the travel times in the congestion model presented in Section II-B are deterministic, $\mathcal{D}^t$ and$\mathcal{C}^t$ do not need to be estimated.

B. Experimental Design

We conduct two experiments in this study. In the first experiment, we compare the performance of the proposed MPC Algorithm 1 with and without the intermodality feature. In the second experiment, we compare the time-variant model from Section II-A to the time-invariant model from [7] to understand how non-stationarity of demand and public transit schedules affect the system operation.

I-AMoD VS AMoD: In the first experiment, we simulate the Manhattan transportation network on March 1st, 2012 from 7pm to 8pm. During this time, the data set comprises 20,000 requests. We use Algorithm 1 to coordinate the I-AMoD fleet with the subway network to serve these requests. As a benchmark, we simulate an AMoD system operating in isolation. Specifically, we use an AMoD-MPC scheme consisting of Algorithm 1 with the subway disabled. Both algorithms control a fleet of 5000 vehicles, and have prediction and optimization horizons of 36 and 40 minutes, respectively. We chose a 40 minute length for the optimization horizon because 99% of the trips in the taxi data set are at most 40 minutes long. Comparing these results, we highlight the significance of considering intermodality in Algorithm 1.

To study the impact of congestion on the system's performance, we run simulations for various levels of exogenous congestion. We quantify exogenous congestion as a percentage share of road capacity, that is the ratio of exogenous vehicles on a road related to its nominal capacity. We consider a service fleet with 5000 vehicles, which is significantly smaller than the amount of vehicles in Manhattan. This way, the approximation of congestion being exogenous is in order, and, therefore, we can remove constraints (4) for Experiment 1 as they are never active.

We evaluate the performance of the algorithms with a combination of the average travel time of the customers and the operational cost incurred by the transportation system operator. Table I summarizes the cost parameters used in our case study and their sources. With these costs, for an arc $(i,j) \in \mathcal{A}$ with travel time t_{ij} and length ℓ_{ij} we have $\rho^t_{ij} = V_{\mathrm{T}} \cdot t_{ij}$ for customer cost and $\rho^o_{ij} = V_{\mathrm{D,R}} \cdot \ell_{ij}$ and $\rho^o_{ij} = V_{\mathrm{D,P}} \cdot \ell_{ij}$ for vehicle and subway arcs respectively. We penalize unserviced requests with a fee of $P = 50\,\mathrm{USD}$, corresponding to a monetary compensation for a loss of about two hours.

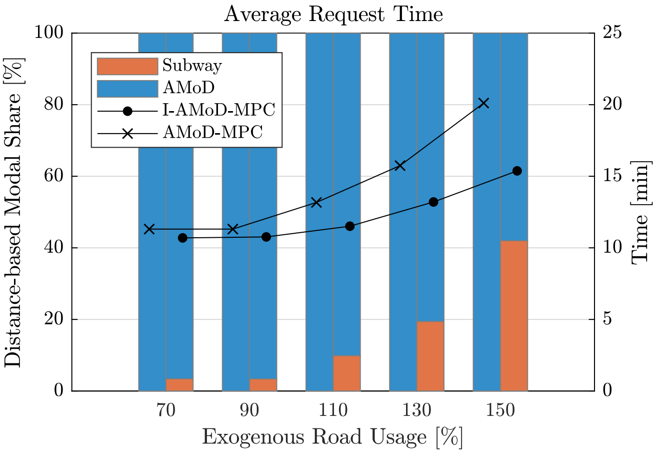

Fig. 4. Comparison of the I-AMoD-MPC and the AMoD-MPC controller. For each level of exogenous congestion, the colored bars specify the distance-based modal share (AMoD or subway) utilized by the AMoD-MPC (left) and the I-AMoD-MPC (right). The black lines denote the average travel time as a function of congestion for both algorithms.

Time-variant VS Time-invariant: To allow for a fair comparison, we make several adjustments to the demand and the time-variant model. First, we scale the original demand by a factor of six to account for the actual number of ride-hailing requests in NYC in 2017 [33]. To avoid a bias while comparing a transient to a steady state modeling approach, we modify the time-variant model as follows: First, since in [7] the algorithm can choose the size of the fleet, we consider $\mathcal{D}_{\text{final}}$ and $\mathcal{D}_{\text{initial}}$ to be decision variables. Second, since the distribution of vehicles in [7] is time-invariant, we include the constraint $\mathcal{D}_{\text{final}} = \mathcal{D}_{\text{initial}}$. Similarly to Experiment 1, we consider various levels of exogenous congestion ranging from 70% to 150% of the nominal road capacity. We optimize over the entire horizon of one hour and analyze the fractional solution $(f^\star_0, f^\star_m)$.

C. Experiment 1: MPC Algorithm Implementation

Fig. 4 shows the results of the comparison between the I-AMoD-MPC and the AMoD-MPC scheme. As can be seen, the total trip times increase when the level of exogenous congestion increases. Even for low levels of congestion, the I-AMoD-MPC is able to decrease the average request time compared to the AMoD-MPC. At high congestion the rebalancing of the vehicles becomes difficult, resulting the AMoD-MPC performing significantly worse than its intermodal counterpart. Indeed, the I-AMoD-MPC suffers less from the increased congestion by allocating more demand to public transit. Overall, the I-AMoD-MPC is able to decrease the total trip time by up to 25%.

We also conducted a system sensitivity analysis to the fleet size. Specifically, we ran the experiment with 130% exogenous road usage for fleet sizes of 4500 and 5500 vehicles. In the former case, the subway share increased from 19% to 27% to compensate for the smaller fleet size,

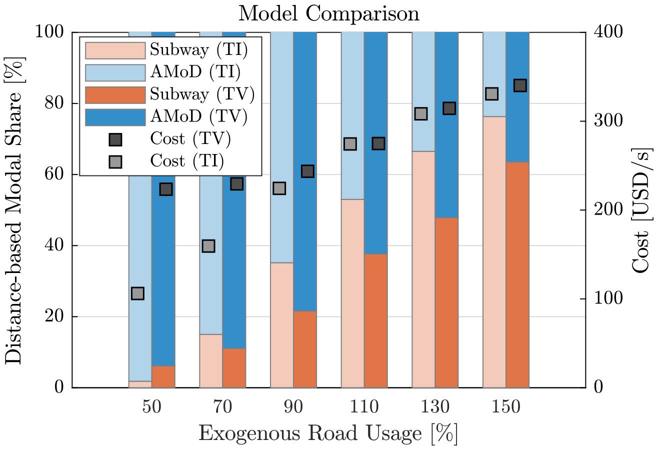

Fig. 5. Comparison between the time-invariant (TI) and the time-variant (TV) I-AMoD system

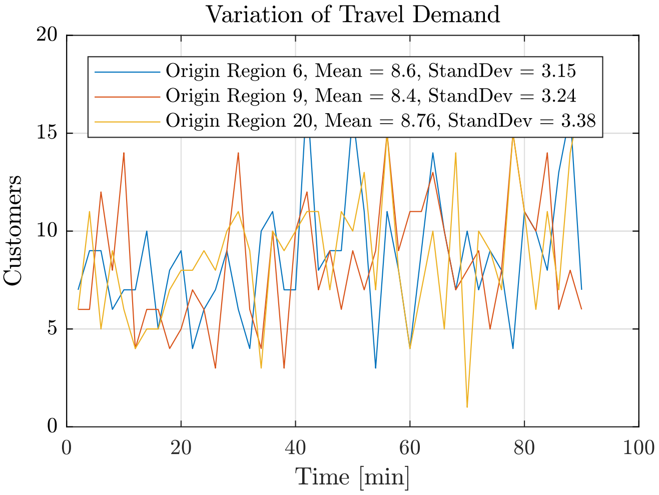

Fig. 6. Temporal variation of the travel demand for different origin regions.

and only increased the average total trip time increased by 48 seconds. In the latter case, the increased fleet size only improves average trip time by 20 seconds. This suggests that cooperation with the subway network can reduce the number of ridehailing vehicles needed.

D. Experiment 2: Model Comparison

Fig. 5 shows the distance-based modal share and the societal cost (travel time and operational cost) of the time-invariant and time-variant modeling approaches for various levels of exogenous congestion. Both models show a monotone behavior: As the exogenous road usage increases, the share of trips served using the subway increases. The time-variant model, however, always has a higher projected cost and lower subway share. This difference is largely explained by two factors. First, the time-invariant model assumes a time-invariant demand. This is only true to some extent. Although the total quantity of requests is relatively stable over the 2 hour time horizon, there are still considerable local spatial and temporal fluctuation in demand. This can be seen in Fig. 6 which shows the the number of requests of a given origin region over time. Second, due to the clustering, the time-variant model typically overestimates travel times due to its lower geographical resolution. This overestimation of travel time, however, makes the model robust to inaccuracies in travel time and congestion models, which is valuable in a time-variant setting where missing the subway can significantly increase the trip time.

E. Discussion

In Experiment 1, the linear optimization problem (12) was solved on a lab workstation (Intel Xeon Gold 6136 CPU, 128 GB RAM) using Gurobi 8.0.1 in less than 10 min. This, however, is not fast enough for real-time applications as a solution is needed once every $\tau = 2$ min. Using computer clusters dedicated to logistics application with distributed optimization schemes such as dual decomposition or alternating direction method of multipliers (ADMM) may enable a real-time implementation of Algorithm 1.

From Experiment 2 we see that low geographical resolution in the time-variant model leads to discrepancies in system cost with the time-invariant model, as shown in Fig. 5. This suggests that a higher resolution should be used for low congestion situations. However, a lower geographical resolution tends to overestimates of travel times and thus provides routes that are robust to inaccuracies in the congestion model.

V. Conclusion

We developed a time-variant network flow based optimization model and leveraged it to devise a MPC algorithm to operate an I-AMoD system that jointly coordinates an AMoD fleet and public transit to service travel demands. We compared the I-AMoD-MPC scheme to a pure AMoD-MPC scheme and showed intermodality to significantly improve service quality. Additionally, we compared the time-variant I-AMoD model with a time-invariant framework which revealed that both high geographical and temporal resolution are needed to obtain high quality solutions.

This work opens the field for several research directions. First of all, the development of a low-level controller would allow for intelligent routing within geographically clustered regions. Furthermore, incorporating fairness for customers is paramount to applications since the current model only aims to optimize the average quality of service. Finally, we would like the MPC algorithm to explicitly account for stochastic effects such as demand fluctuation, congestion deviations and public transit delays.

Acknowledgments

We would like to thank Mr. Nicolas Lanzetti and Dr. Kiril Solovey for their precious help and the fruitful discussions. Additionally, we thank Dr. I. New for her assistance with the proofreading and useful advice. 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.

References

[1] J. I. Levy, J. J. Buonocore, and K. Von Stackelberg, "Evaluation of the public health impacts of traffic congestion: a health risk assessment," *Environmental Health*, vol. 9, no. 1, p. 65, 2010.

[2] UN DESA. (2018) 68% of the world population projected to live in urban areas by 2050. retrieved from:. Available at http://urs-srv-eprints.u-strasbg.fr/337/01/AUBRIOT_S%C3%A9bastien_2008.pdf.

[3] W. Hu. (2017) Your Uber car creates congestion. should you pay a fee to ride? The New York Times. Available online.

[4] P. Berger. (2018) Mta blames uber for decline in new york city subway, bus ridership. The Wall Street Journal. Available online.

[5] M. Pavone, S. L. Smith, E. Frazzoli, and D. Rus, "Load balancing for Mobility-on-Demand systems," in *Robotics: Science and Systems*, 2011.

[6] 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, in Press. Extended Version, Available at https://arxiv.org/abs/1804.11278.

[7] M. Salazar, N. Lanzetti, F. Rossi, M. Schiffer, and M. Pavone, "Intermodal autonomous mobility-on-demand," *IEEE Transactions on Intelligent Transportation Systems*, 2019, submitted. [Online]. Available: ../wp-content/papercite-data/pdf/Salazar.ea.T-ITS19.pdf

[8] 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.

[9] 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.

[10] M. Maciejewski, J. Bischoff, S. Hörl, and K. Nagel, "Towards a testbed for dynamic vehicle routing algorithms," in *Int. Conf. on Practical Applications of Agents and Multi-Agent Systems - Workshop on the application of agents to passenger transport (PAAMS-TAAPS)*, 2017.

[11] 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.

[12] M. Pavone, K. Treleaven, and E. Frazzoli, "Fundamental performance limits and efficient policies for Transportation-On-Demand systems," in *Proc. IEEE Conf. on Decision and Control*, 2010.

[13] K. Spieser, 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.

[14] 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.

[15] 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, submitted.

[16] 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, submitted.

[17] F. Rossi, 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.

[18] R. Iglesias, F. Rossi, K. Wang, D. Hallac, J. Leskovec, and M. Pavone, "Data-driven model predictive control of autonomous mobility-on-demand systems," in *Proc. IEEE Conf. on Robotics and Automation*, 2018.

[19] M. Tsao, 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.

[20] M. Tsao, D. Milojevic, C. Ruch, M. Salazar, E. Frazzoli, and M. Pavone, "Model predictive control of ride-sharing autonomous mobility on demand systems," in *Proc. IEEE Conf. on Robotics and Automation*, 2019, in Press.

[21] G. Gentile and K. Noekel, Eds., *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*. Springer New York, 2016.

[22] J. Bischoff, I. Kaddoura, M. Maciejewski, and K. Nagel, "Re-defining the role of public transport in a world of shared autonomous vehicles," in *Symposium of the European Association for Research in Transportation (hEART)*, 2017.

[23] A. Vakayil, W. Gruel, and S. Samaranayake, "Integrating shared-vehicle Mobility-on-Demand systems with public transit," in *Annual Meeting of the Transportation Research Board*, 2017.

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

[25] J. G. Wardrop, "Some theoretical aspects of road traffic research," *Proc. of the Institution of Civil Engineers*, vol. 1, no. 3, pp. 325–362, 1952.

[26] 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.

[27] F. Rossi, R. Iglesias, R. Zhang, and M. Pavone. (2017) Congestion-aware randomized routing in autonomous mobility-on-demand systems. Extended version Available at https://asl.stanford.edu/wp-content/papercite-data/pdf/Rossi.Iglesias.Zhang.Pavone.CDC17.pdf.

[28] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.

[29] New York City Open Data. (2018) New York City Subway Lines. Available at https://data.cityofnewyork.us/Transportation/Subway-Lines/3qz8-muuul.

[30] U.S. Dept. of Transportation, "Revised departmental guidance on valuation of travel time in economic analysis," Tech. Rep., 2015.

[31] J. Neff and M. Dickens, "2016 public transportation fact book," American Public Transportation Association, Tech. Rep., 2017.

[32] Bureau of Transportation Statistics, "National transportation statistics." U.S. Dept. of Transportation, Tech. Rep., 2016.

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

Appendix

A. Intermodal Structure of the Time-expanded Graph

For further clarity we would like to elaborate on the categorization of arcs in $\mathscr{A}$. Arcs between vertices can be categorized into two categories: intra-regional and inter-regional arcs. Arcs that connect different vertices in one region are intra-regional arcs. They model waiting time, mode switching, customer pickup and delivery. The intra-regional arcs of a region are represented in Fig. 2. The orange arrows represent arcs in $\mathscr{A}_{\text{cus}}$. They model the the fact that customers can wait inside the region and at public transit stations. The black solid arrows represent arcs in $\mathscr{A}_{\text{veh}}$. Such an arrow from a road vertex to a walking vertex of a region models the drop off of a customer in the region, which takes one time-step τ. Equivalently, here are black solid arrows from the road vertex to the subway vertex for a drop off directly at the public transit station. The pickup of the customer is represented as well by black solid arrows but from walking or subway vertices to the road vertex. The black dotted arrows represent arcs in $\mathscr{A}_{\text{reb}}$. They model the fact that once a customer is dropped off by a vehicle, the vehicle can directly leave the region via a road vertex or enable the pickup of a customer in the same region.

Arcs that connect vertices of different regions are inter-regional arcs. They either connect vertices of the road network or subway vertices. The connection between road vertices is given by the road arc set $\mathscr{A}_{\text{R}}$, which is a subset of $\mathscr{A}_{\text{veh}}$. $\mathscr{A}_{\text{R}}$ results from the topology of the road network and the travel time between regions, depending on the levels of exogenous congestion. The inter-regional connections between public transit station represent the topology of the public transit network and schedule. If some public transit stations are geographically close but not connected by a public transit line they are connected by an arc that represents the time it takes to walk in between them, to allow the switching of lines.