Abstract—Mobility-on-demand (MoD) systems are revolutionizing urban transit with the introduction of ride-sharing. Such systems have the potential to reduce vehicle congestion and improve accessibility of a city’s transportation infrastructure. Recently developed algorithms can compute routes for vehicles in real-time for a city-scale volume of requests while allowing vehicles to carry multiple passengers at the same time. However, these algorithms focus on optimizing the performance for a given fleet of vehicles and do not tell us how many vehicles are needed to service all the requests. In this paper, we present an offline method to optimize the vehicle distributions and fleet sizes on historical demand data for MoD systems that allow passengers to share vehicles. We present an algorithm to determine how many vehicles are needed, where they should be initialized, and how they should be routed to service all the travel demand for a given period of time. Evaluation using 23,529,740 historical taxi requests from one month in Manhattan shows that on average 2864 four passenger vehicles are needed to service all of the taxi demand in a day with an average added travel delay of 2.8 mins.

I. INTRODUCTION

Autonomous vehicles and transportation network companies are revolutionizing personal mobility by making transportation available anywhere at anytime. Mobility-on-demand (MoD) has the potential to provide faster and more efficient transportation using fleets of coordinated autonomous vehicles. State of the art algorithms are able to efficiently manage fleets of vehicles to service large volumes of requests as is needed in dense urban areas.

Ride-sharing, where more than one passenger is able to use the same vehicle at the same time, is improving the efficiency of vehicle fleets by allowing less vehicles to service more requests. Services such as UberPool and Lyft Line have demonstrated how effective ride-sharing can be at extending our means of transportation within a city. Global vehicle dispatching algorithms have been developed that make use of ride-sharing by considering batches of requests, and optimizing routes for a set of vehicles to service these requests. These algorithms can ensure that passengers do not need to wait too long to be picked up and that by sharing a ride, the passenger will not go too far out of their way. However, these approaches cannot guarantee that all requests are serviced and do not tell us how many vehicles are needed. Being able to give insight to a fleet operator as to how many vehicles they need is incredibly important because it can help them plan for deployments in new areas, inform them which areas are being underserved, and improve the efficiency of their fleet.

In what follows, we address the problem of determining how many vehicles are needed and where they should be for a MoD fleet to service all the taxi demand at city-scale with a maximum waiting time and maximum incurred delay while allowing multiple requests to be serviced by the same vehicle. Our method runs offline and can inform a fleet operator of the size and distribution of their fleet needed to satisfy historical demand. We show that optimizing the number of vehicles and their distribution, we can significantly improve the efficiency of a MoD fleet.

A. Related Work

Early work for MoD systems focused on developing algorithmic techniques for fleets with single occupancy vehicles [1]–[6]. However, with the explosion of the sharing economy [7], [8], there is great potential for ride-sharing to make our fleets more affordable and efficient.

Recently, algorithms for large-scale ride-sharing have been developed. The work in [9] showed that 80% of rides in Manhattan could be pairwise shared while increasing the
travel time by only a couple minutes. They later extended the 
analysis to multiple cities [10]. Alonso-Mora et al. developed 
a scalable approach to allow more than two passengers to 
share a vehicle by finding an optimal assignment of requests 
to a given fleet of vehicles [11]. The efficiency of the fleet 
improved in [12] by proactively routing the vehicles through 
areas of high demand. However, these approaches did not 
determine how many vehicles would be needed to service 
all the travel demand.

There has been some research focusing on fleet sizing 
but either analyzes how different fleet sizes will perform to 
satisfy travel demand [13] or use a fixed start and end point 
model [14]. Recently, major breakthroughs have been made 
using offline methods to determine the fleet size needed to 
satisfy historical demand data. Vazifeh et al. [15] provides 
a method to optimally solve the minimum fleet problem 
online on historical demand data, but assumes a vehicle 
can only transport one passenger at a time and cannot be 
easily extended to ride-sharing. Čap and Alonso-Mora [16] 
used multi-objective analysis to determine the required fleet 
size for a set of requests while allowing multiple requests to 
share a ride, but the method was not scalable and they only 
presented experiments using 1 minute of request data.

B. Method Overview

The method for optimizing the fleet size and vehicle 
distribution is split into multiple steps. First we select a 
set of starting locations we call vehicle deposits such that 
every node in the road network is reachable from at least 
one deposit within given amount of time. These deposits are 
computed once offline and remain constant. We then iterate 
over the set of requests in batches backwards in request time.

The batch size is fixed for all iterations and we only iterate 
over the batches once. One can think of the set of requests 
as a stack of the remaining requests to satisfy and at each 
iteration we are popping off the latest batch.

For each batch, we compute a set of travel schedules to 
pick up and drop off requests in that batch using the method 
from [11]. We assume vehicles are located at the vehicle 
deposits when computing these schedules.

While iterating over batches of requests, we maintain a 
set of initial travel schedules. These are the first schedules 
vehicles would execute after initializing. For the first iter-
ation, this set is empty. For each iteration, we determine 
how vehicles can transition between schedules computed 
for the current batch of requests to the set of initial travel 
schedules. We then select a set of schedules and transitions 
from those computed that minimizes a cost function, satisfies 
certain constraints, and services all travel requests from the 
batch. Any initial travel schedule that receives an incoming 
transition is then removed from the set and all selected 
schedules are added to the set.

After iterating over all requests, we compute a maximum 
bipartite matching between travel schedules that have no 
outgoing transitions and our set of initial travel schedules 
using more lenient transition constraints. Schedules in the 
initial travel schedule set that receive incoming transitions 
from this matching are removed. Afterwards, this set of 
initial travel schedules will tell us how many vehicles we 
need and where they should be to service all the travel 
demand.

II. PRELIMINARIES

A. Definitions

We assume that the vehicles travel along a road-network, 
\( G = (N, E) \), represented as a directed graph, and that we 
have a function \( \tau(n_i, n_j) \) for \( n_i, n_j \in N \) that gives the 
shortest travel times between nodes in the graph.

We consider a set of travel requests, \( \mathcal{R} \), and a set of 
vehicle deposits, \( \mathcal{D} \). A travel request is a tuple \( r = (p_r, d_r, t_r^{\text{req}}) \), 
where \( p_r \in N \) is the pickup location, \( d_r \in N \) is the dropoff 
location, and \( t_r^{\text{req}} \) is the time the request was made. \( \mathcal{R} \) 
is sorted in ascending order by request time. \( \mathcal{R} \) is iterated over 
in batches of \( n \) requests backwards in time. That means the 
first batch will consist of requests from \( r_{\mathcal{R}} \) to \( r_{\mathcal{R}} \).

A vehicle deposit is a location on the graph, \( D \subseteq N \), where vehicles are initialized. Vehicle deposits can be 
thought of as starting locations for vehicles before they have 
been assigned any requests.

For the formulation we consider a travel request to be 
satisfied if 1) the waiting time, \( \omega_r \), given by the difference 
between the pickup time, \( t_r^p \), and the request time, \( t_r^{\text{req}} \) is less 
than a specified maximum waiting time, \( T_{\text{wait}} \) and 2) the total 
travel delay for the request given by \( \delta_r = t_r^d - t_r^p \) is less than a 
specified maximum travel delay, \( T_{\text{delay}} \), where \( t_r^d \) is the time 
when the request is dropped off and \( t_r^d = t_r^{\text{req}} + \tau(p_r, d_r) \) is 
the earliest possible time the destination could be reached.

For the approach, we compute a set of travel schedules 
and schedule transitions that vehicles will follow to satisfy 
the travel requests in \( \mathcal{R} \). A travel schedule, \( S = \{ (l \in N, t \in \mathbb{R}_+) \ldots \} \) is a sequence of locations a vehicle will visit and at 
what times to service a set of requests denoted as \( R(S) \subseteq \mathcal{R} \). 
Each \( (l, t) \in S \) corresponds to either a pick up or drop 
off. For convenience, let’s call \( \ell_S(k) \) and \( t_S(k) \) the location 
and time of the \( k \)th event in the schedule respectively. We 
assume that the vehicle takes the shortest path between those 
locations. For a schedule to be valid, all the travel requests 
in the schedule must be satisfied given the maximum waiting 
time and maximum travel delay constraints and the number 
of passengers in the vehicle must not surpass the vehicle’s 
capacity at any given time. We define the cost of each 
schedule as the sum of the travel delays for the requests 
associated with the schedule \( \delta(S) = \sum_{r \in R(S)} \delta_r \).

A schedule transition is a pair of schedules, \( (S_i, S_j) \), such 
that a vehicle is able to satisfy \( S_i \) then \( S_j \) without idling for 
longer than a given maximum vehicle idle time, \( T_{\text{idle}} \). Let’s 
define the transition time and idle time between schedules 
as \( \xi(S_i, S_j) = \tau(\ell_{S_i}, |S_i|), \ell_{S_j}(1)) \) and \( \theta(S_i, S_j) = t_{S_j}(1) - t_{S_i}(|S_i|) - \xi(S_i, S_j) \) respectively.

B. Selecting Vehicle Deposits

Due to the maximum waiting time and maximum delay 
constraints for travel requests, it is not possible to guarantee a 
prescribed service rate for an arbitrary set of vehicle deposits.
For instance, if the travel time from the closest vehicle deposit to a request’s pick up location is larger than the maximum waiting time, that request may be impossible to service. Therefore, in order to provide a guaranteed service rate, we must intelligently select the locations for the vehicle deposits.

Using the road-network, \( G = (N,A) \), we select a set \( D \subseteq \mathbb{N} \) as vehicle deposit locations such that \( \forall n \in \mathbb{N}, \exists d \in D \) with \( \tau(d,n) \leq T_{\text{depos}} \), where \( 0 \leq T_{\text{depos}} \leq T_{\text{wait}} \). To reduce the computational overhead for computing schedules, we select the minimum number of vehicle deposits needed. We can do this by solving an integer linear program. First, we define a reachability matrix as \( H_{ij} = 1 \) for nodes \( n_i, n_j \in \mathbb{N} \) if \( \tau(n_i,n_j) \leq T_{\text{depos}} \) and 0 otherwise. This matrix describes which nodes are reachable from a given node within a specified amount of time.

We also define a set of binary variables \( x \in \{0,1\}^{\vert N \vert} \) where \( x_i = 1 \) if \( n_i \) is used as a deposit location and 0 otherwise. We can now solve an integer linear program (ILP) to determine the minimum number of nodes to use as vehicle deposits such that the reachability constraint is satisfied:

\[
\begin{align*}
\min & \quad \sum_{i=1}^{\vert N \vert} x_i \\
\text{s.t.} & \quad \sum_{i=1}^{\vert N \vert} x_i \cdot H_{ij} \geq 1 \quad \forall j \in \{1,\ldots,\vert N \vert \}
\end{align*}
\]

Eq. (2) guarantees that every node in \( N \) is reachable within \( T_{\text{depos}} \) travel time from at least one vehicle deposit. Now we can define the set of vehicle deposits as \( D = \{n_i : 1 \leq i \leq \vert N \vert \wedge x_i = 1\} \).

### III. FLEET OPTIMIZATION

In this section we describe how schedules and schedule transitions are selected for each batch using an integer linear program (ILP) and how the set of initial travel schedules is updated after each iteration.

#### A. Schedule Chaining Overview

We compute the schedule chains using an iterative batch process working backwards in time. At each iteration we select a set of schedules and schedule transitions that minimizes the cost function and satisfies the constraint set. After each iteration, the selected schedules and schedule transitions are fixed and must remain selected. Any schedule that is selected that does not have an incoming transition is called a start as it requires a new vehicle. At each iteration we will maintain a set of starts that could be used for future transitions. This reduces the computational overhead of the approach.

At the \( k \)th iteration, we are given a set of requests, \( R_k \subseteq \mathcal{R} \), ordered by request time. We can consider \( R_k \) the remaining requests left to service after \( k-1 \) batches have been processed. For the first iteration, \( k = 1 \), and we have \( R_1 = \mathcal{R} \). The batch of requests for the \( k \)th iteration, \( R_k^{\text{latest}} \subseteq R_k \), are the \( n \) latest requests from \( R_k \), that is the \( n \) requests in \( R_k \) with the latest request times. We then compute schedules using the requests in \( R_k \) that service requests in \( R_k^{\text{latest}} \) using the method from [11]. We limit the number of schedules by only considering requests in \( R_k \) that occurred within some time, \( T_{\text{req}} \) since the requests in \( R_k^{\text{latest}} \) were made. Note that these schedules will contain all the requests in \( R_k^{\text{latest}} \) but may also contain other requests from \( R_k \). We can think of this set of schedules as a set of candidate schedules that service requests in \( R_k^{\text{latest}} \). We will call this set of candidate schedules, \( S_k \) where \( S_k,i \) is the \( i \)th schedule in that set. We will define the set of associated requests for the set of schedules, \( S_k \), as \( R_k^{\text{assoc}} = \bigcup_{i=1}^{\vert S_k \vert} R(S_k,i) \).

At each iteration, we want to select a set of schedules, \( S_k^* \), that would require new vehicles (i.e. schedules that have no incoming transitions). We select these schedules from the computed schedules for the current iteration, \( S_k \), and schedules selected from the previous iteration, \( S_{k-1}^{\text{assoc}} \). At the first iteration, \( S_{k-1}^{\text{assoc}} = S_0^{\text{assoc}} = \emptyset \). In conjunction with selecting initial schedules, we select a set of schedule transitions out of candidate transitions from \( S_k \) to \( S_{k-1}^{\text{assoc}} \). The transitions and schedules, \( S_k^* \), are solved for at the same time. Note that not all schedules in \( S_k^{\text{assoc}} \) need to be in \( S_k^* \). If a transition is computed between some \( u_k \in S_k^* \) and \( v_{k-1} \in S_{k-1}^{\text{assoc}} \), then \( v_{k-1} \) will not be in \( S_k^* \) since it would not require a new vehicle. Since we are iterating through the request set backwards in time, at the final iteration, we will have a set of schedules where vehicles will begin their routes. This will give us an initial distribution of vehicles and by iterating through the transitions, will provide the temporal distribution.

#### B. Integer Linear Program Formulation

We formulate this schedule selection problem as an ILP. We define a set of binary variables \( s \in \{0,1\}^{\vert S_k \vert} \) where \( s_i = 1 \) if schedule \( i \) is selected from \( S_k \) and zero otherwise.

We also define a set of binary transition variables, \( \epsilon \in \{0,1\}^{\vert S_k \vert \times \vert S_{k-1}^{\text{assoc}} \} \) where \( \epsilon_{ij} = 1 \) if schedule \( i \) in \( S_k \) should transition to schedule \( j \) in \( S_{k-1}^{\text{assoc}} \) and zero otherwise. Lastly we define two more sets of binary variables, \( \chi \in \{0,1\}^{\vert S_k \vert} \) and \( \eta \in \{0,1\}^{\vert S_{k-1}^{\text{assoc}} \} \) where \( \chi_i = 1 \) if the \( i \)th request in \( R_k^{\text{latest}} \) was not used and \( \eta_i = 1 \) if the \( i \)th schedule in \( S_{k-1}^{\text{assoc}} \) was not used. With these variables, \( \mathcal{Y} = (s,\epsilon,\eta,\chi) \), we can define the cost function as:

\[
C(\mathcal{Y}) = \sum_{i=1}^{\vert S_k \vert} \left( \delta(S_{k,i}) \cdot s_i + \sum_{j=1}^{\vert S_{k-1}^{\text{assoc}} \vert} \xi(S_{k,i};S_{k-1,j}) \cdot \epsilon_{ij} \right) \\
+ K_r \sum_{i=1}^{\vert S_k \vert} \chi_i + K_s \sum_{i=1}^{\vert S_{k-1}^{\text{assoc}} \vert} \eta_i
\]

This cost combines the sum of the delays for the selected schedules from \( S_k \), the transition times for schedule transitions selected from \( S_k \) to \( S_{k-1}^{\text{assoc}} \), and additive costs for ignoring requests in \( R_k^{\text{assoc}} \) and schedules in \( S_{k-1}^{\text{assoc}} \). In Eq. (3), \( K_r \) and \( K_s \) are constant cost values for ignoring requests and schedules respectively.

1) Requests are serviced at most once: For a schedule selection to be valid, we must ensure that each request in \( R_k^{\text{assoc}} \) is serviced at most once. This is described in the
constraint:
\[ \chi_i + \sum_{j \in T^{S_{r}}_{i}} s_j = 1 \quad \forall i \in \{1, \ldots, |R_{assoc}^k|\} \quad (4) \]

In Eq. (4), the set \( T^{S_{r}}_{i} \) corresponds to the indices of schedules in \( S_k \) that service the \( i \)th request in \( R_{assoc}^k \).

2) Minimal acceptable service rate: We want to select schedules that will guarantee at least a minimum acceptable service rate is achieved. This is accomplished by ensuring that at least \( \beta \cdot n \) are serviced out of the \( n \) latest requests, \( R_{latest}^k \) where \( 0 < \beta \leq 1 \) is a quality-of-service parameter. This constraint is formulated as:
\[ \sum_{i=1}^{|R_{assoc}^k|} \chi_i \leq \beta \cdot n \quad (5) \]

3) Maximum idle time: To ensure that the vehicles do not idle for more than a specified amount of time, \( T_{idle} \), between transitions, we must constrain transition variables. Though we already take into account the transition times into the cost function, directly constraining the amount of vehicle idling allows us to provide guarantees for the maximum idle times for vehicles and reduces the size of the problem. This constraint is formulated as:
\[ \epsilon_{ij} \cdot \theta(S_{k,i}, S_{k-1,j}) \leq T_{idle} \quad (6) \]

\[ \forall i \in \{1, \ldots, |S_k|\} \text{ and } \forall j \in \{1, \ldots, |S_k^*|\} \]

4) Schedules start on time: To guarantee that requests in schedules selected in previous iterations are still satisfied, we must only select transitions such that the previously selected schedules, \( S_{k-1}^* \), start on time. This constraint over the transition variables is formulated as:
\[ \epsilon_{ij} \cdot (t(S_k,i,(|S_k|,i)) + \xi(S_k,i,(|S_k|,i,j)) \leq t_{S_{k-1,j}} \quad (7) \]

\[ \forall i \in \{1, \ldots, |S_k|\} \text{ and } \forall j \in \{1, \ldots, |S_k^*|\} \]

This constraint ensures that the following schedule will be reached in time such that the waiting time and delay constraints are satisfied. This constraint also guarantees that all other schedules in the chain are also satisfied.

5) Schedule flow constraints: Also, to enable valid transitions, we must apply flow constraints to the transition variables. There can only be an outgoing transition from schedule \( i \) if the \( i \)th schedule is selected. This is described as:
\[ |S_{k-1}^*| \sum_{j=1}^{S_{k-1}^*} \epsilon_{ij} \leq s_i \quad \forall i \in \{1, \ldots, |S_k|\} \quad (8) \]

Likewise, we can only have at most one incoming transition to a given schedule in \( S_{k-1}^* \) from \( S_k \). This is described by the constraint:
\[ \eta_j + \sum_{i=1}^{S_{k-1}^*} \epsilon_{ij} = 1 \quad \forall j \in \{1, \ldots, |S_{k-1}^*|\} \quad (9) \]

Note that we have introduced the variable \( \eta_j \) here to indicate if the \( j \)th schedule in \( S_{k-1}^* \) is not selected. This is used in the cost function.

6) Complete ILP description: Combining the cost function from Eq. (3) with the constraints described in this section, we can formulate an ILP that will select a set of schedules and schedule transitions that will minimize the cost function while satisfying service rate, delay, waiting time, and transition constraints for the set of given requests. The full ILP is then formulated as:
\[ \min \quad \mathcal{C}(Y) \quad (10) \]

s.t. constraints (4) – (9)

C. Preparing For Next Iteration

After solving this ILP, we need to gather the selected schedules in preparation for the next iteration. To determine \( S_{k}^* \), we first add all schedules from \( S_k \) for which the corresponding variable \( s_i = 1 \). We also add all schedules from \( S_{k-1}^* \) that do not have any incoming transitions. These are all the schedules which would require a new vehicle to service since they are not going to be transitioned to by another vehicle after finishing its schedule. This set is constructed as:
\[ S_{k}^* = \{ S_{k,i} : s_i = 1, \forall i \in \{1, \ldots, |S_{k,i}|\} \} \]

\[ \cup \{ S_{k-1,j}^* : \eta_j = 1, \forall i \in \{1, \ldots, |S_{k-1,j}|\} \} \quad (11) \]

For the next iteration we also need to remove all the requests from the set of requests that have already been serviced by schedules selected in the current iteration:
\[ R_{k+1} = R_k \setminus \bigcup_{i=1}^{|S_{k}^*|} R(S_{k,i}) \quad (12) \]

Finally, to keep track of all the schedules and associated transitions, we maintain a set of all schedules selected, \( S_k \), and a function \( N : S \rightarrow S \), that maps all schedules to their transitions. In the case that a given schedule \( s \in S \) has no outgoing transitions, \( N(s) = \emptyset \).

Now that the remaining requests have been gathered and schedules have been selected, we can move on to the next iteration. An overview of the schedule chaining algorithm is shown in Algo. 1. In the algorithm we use the short hand functions LatestRequests(\( R_k \)) to represent the \( n \) latest requests from \( R_k \). ComputeSchedules(\( R_{latest}^k \)) to represent all the schedules computed from \( R_{latest}^k \), and SolveScheduleSelectionILP(\( S_k, S_{k-1}^*, R_{latest}^k, R_{assoc}^k \)) to represent solving Eq. (10) and extracting the schedules selected from the optimization variables.

IV. Long Term Rebalancing

After computing the starting schedules using the schedule chaining algorithm, there may be more starting schedules selected than necessary due to the constraint on the maximum idling time for a vehicle. We perform what we call long term rebalancing, which relaxes the constraint on the maximum idling time after the last iteration of the schedule chaining algorithm to reduce the total number of vehicles needed in the fleet. Long term rebalancing computes a matching from schedules which have no outgoing transitions to schedules that have no incoming transitions using a larger maximum
Algorithm 1 Overview of the Schedule Chaining Algorithm
1: \( S \leftarrow \{ \} \)
2: \( S_0 \leftarrow \{ \} \)
3: \( R_1 \leftarrow \text{AllRequests}() \)
4: \( k \leftarrow 1 \)
5: while \( |R_k| > 0 \) do
6: \( R_k^{\text{latest}} \leftarrow \text{LatestRequests}(R_k) \)
7: \( S_k \leftarrow \text{ComputeSchedules}(R_k^{\text{latest}}) \)
8: \( R_k^{\text{assoc}} = \bigcup_{i=1}^{|S_k|} R(S_k,i) \)
9: \( S_k^{*} \leftarrow \text{ScheduleSelectionILP}(S_k, S_{k-1}^{*}, R_k^{\text{latest}}, R_k^{\text{assoc}}) \)
10: \( R_{k+1} \leftarrow R_k \setminus \bigcup_{i=1}^{|S_k|} R(S_k^{*}, i) \)
11: \( S \leftarrow S \cup S_k^{*} \)
12: for \( i = 1 \) to \( |S_k| \) do
13: for \( j = 1 \) to \( |S_{k-1}^{*}| \) do
14: if \( s_i = 1 \) and \( \epsilon_{ij} = 1 \) then
15: \( N(S_{k,i}) \leftarrow S_{k-1,j}^{*} \)
16: end if
17: end for
18: end for
19: \( k \leftarrow k + 1 \)
20: end while

idling time, denoted as \( T_{\text{reb}} \), while ensuring the delay and waiting time constraints for the future schedules remain satisfied.

We define the two sets for the matching problem, the set of schedules with no outgoing transitions, \( A = \{ s : N(s) = \emptyset, \forall s \in S \} \), and the set of schedules without any incoming transitions, \( B = \{ s : \exists u \text{ such that } N(u) = s, \forall s \in S \} \). To minimize the total number of vehicles needed, we will perform a maximum cardinality bipartite matching between sets \( A \) and \( B \). We will do so using an ILP.

To formulate this ILP, we define a set of binary variables, \( \epsilon = \{(0,1)^{|A| \times |B|}\} \) which represent if schedule \( A_i \) should transition to schedule \( B_j \). With these variables, we can maximize the number of transitions to minimize the total number of vehicles needed in the fleet. The utility function is given as:

\[
J(E) = \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} \epsilon_{ij} \quad (13)
\]

We also need to constrain the matching so that the idling and waiting times for the schedules are respected. These are the same constraints as in Eq. (6) and (7), however a larger vehicle idling time, \( T_{\text{reb}} \), is used:

\[
\epsilon_{ij} \cdot \theta(A_i, B_j) \leq T_{\text{reb}} \quad (14)
\]

\[
\epsilon_{ij} \cdot \left( t_{A_i}(|A_i|) + \xi(A_i, B_j) \right) \leq t_{B_j}(1) \quad (15)
\]

\[\forall i \in \{1, \ldots, |A|\} \text{ and } \forall j \in \{1, \ldots, |B|\}\]

Since we are computing a matching, we need to apply flow constraints to the transition variables to guarantee that there will be at most one outgoing transition for any schedule in \( A \) and at most one incoming transition to any schedule in \( B \):

\[
\sum_{i=1}^{|A|} \epsilon_{ij} \leq 1 \quad \forall j \in \{1, \ldots, |B|\} \quad (16)
\]

\[
\sum_{j=1}^{|B|} \epsilon_{ij} \leq 1 \quad \forall i \in \{1, \ldots, |A|\} \quad (17)
\]

Now we can combine these constraints along with the cost function in Eq. (13) to formulate the ILP for long term rebalancing:

\[
\max \ E \quad J(E) \quad (18)
\]

s.t. constraints (14) – (17)

V. Evaluation

We evaluate the proposed algorithm using historical taxi request data from Manhattan [17] and compare the performance to the actual efficiency of the taxi fleet from the data.

A. Experimental Setup

For the experiments, we used one month of historical taxi data from 00:00 on May 1st to 23:59 on May 31st, 2013. The data contains the origin, destination, pick up time and drop off time for all taxi requests in Manhattan. From this raw data, we use the reported pick up time as the request time since request time was not provided. The road network we use is extracted from OpenStreetMap [18] and the travel times were queried from Google Maps. The shortest paths and travel times between every pair of nodes in the road network were computed offline. We ran the algorithm independently for each day of the month. A computer with a 3.0 GHz, 36 core (72 thread) processor and 144GB of memory was used to run the experiments and we ran four instances of the algorithm at the same time (18 threads per instance).

We assess the performance of the algorithm using vehicles of capacity two and four. We use a fixed maximum waiting time of \( T_{\text{wait}} = 3 \) minutes and a maximum delay of \( T_{\text{delay}} = 6 \) minutes. The vehicle deposits were computed using a maximum travel time of \( T_{\text{depos}} = 1 \) minute and we use a batch size of \( n = 10 \) requests. The minimum acceptable service rate is \( \beta = 1 \) (i.e. we want to service all requests). The maximum idle time for schedule chaining is \( T_{\text{idle}} = 30 \) seconds and for rebalancing \( T_{\text{reb}} = 24 \) hours.

We compare the fleet sizes produced by the proposed algorithm with the actual fleet sizes used to service the requests.

B. Results

We collect several metrics to access the performance of the proposed algorithm including the travel delay, waiting time, vehicle idle time, passenger load, computational time, and fleet size. The travel delay and waiting time are defined in Sec. II-A. The vehicle idle time is the amount of time a vehicle would have to idle between schedules and is also precisely defined in Sec. II-A. The passenger load is the maximum number of passengers in a vehicle at any time for a given schedule. The computation time includes the
Table I: Performance metrics and fleet sizes for vehicle capacities of two and four passengers compared to the performance of the current fleet of NYC taxis.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity=2</td>
<td>1.7</td>
<td>4.8</td>
<td>1.9</td>
<td>35.4</td>
<td>7.6</td>
<td>3761.5</td>
</tr>
<tr>
<td>Capacity=4</td>
<td>2.8</td>
<td>4.7</td>
<td>3.1</td>
<td>41.9</td>
<td>5.9</td>
<td>2864.0</td>
</tr>
<tr>
<td>Actual</td>
<td>0.0</td>
<td>22.3</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>12237.4</td>
</tr>
</tbody>
</table>

We observe a large reduction in the required fleet size to service all the requests when using the proposed algorithm against the baseline. For a fleet with a two passenger vehicle capacity, on average we need 3761 vehicles and for a four passenger vehicle capacity fleet, we need 2864 vehicles. This is a 69% and 77% reduction in fleet size respectively compared to the baseline. The proposed algorithm also reduces the average vehicle idle time significantly from 22.3 mins to 4.7 mins for capacity two and 4.8 mins for capacity four. This is due to the fleet operating more efficiently.

The reduction in fleet size comes at the cost of an added travel delay, however our data shows that this cost is low. Passengers on average would expect to experience an average travel delay 1.7 mins for a capacity two fleet and 2.8 mins for a capacity four fleet.

We can see that the proposed algorithm is efficiently utilizing the vehicles by inspecting the average passenger load. For capacity two fleets, the average load is 1.9 passengers per schedule (95% seat utilization) and for capacity four fleets, it is 3.1 passengers per schedule (76% seat utilization).

The time it took to run the algorithm to determine how many vehicles are needed, where they should be initialized and how they should move to service all the requests for a given day Fig. 2. In our experiments we observe that fleets of capacity two always have higher fleet sizes and take longer to compute than fleets of capacity four for a given number of requests. For our sample, we observe a linear trend in the fleet size against the number of requests for both vehicle capacities. We see that fleet size grows more rapidly for capacity two than for capacity four. For all days the fleet size remains under 4500 vehicles for capacity two and 3500 for capacity four.

VI. CONCLUSION

In this paper, we presented a method to optimize the vehicle distributions for a mobility-on-demand fleet. We presented an algorithm to determine how many vehicles are needed, where they should be initialized, and how they should move to service all the travel demand for a given period of time while allowing multiple passengers to be serviced by the same vehicle. The algorithm can be used to inform a ride-sharing fleet operator of how vehicles should be distributed to handle the demand for an area.

We demonstrated a significant improvement in the vehicle efficiency of the taxi fleet when using the proposed algorithm as compared to how taxis are currently utilized in Manhattan. On average, we can reduce the fleet size by 69% by allowing up to two passengers per vehicle and by 77% for up to four passengers per vehicle while guaranteeing all travel requests are serviced compared to the baseline. These benefits come at only a small cost to the user with an added travel delay of 1.7 mins and 2.8 mins for vehicle capacities of two and four respectively.

Future work will focus on determining a lower bound on the fleet size so we have a range to inform fleet operators. We also plan to investigate how this method can be extended for online use for real-time fleet management.
REFERENCES


