A Mechanism for Organizing Last-Mile Service Using Non-Dedicated Fleet

Shih-Fen Cheng
School of Information Systems
Singapore Management University
Republic of Singapore
sfcheng@smu.edu.sg

Duc Thien Nguyen
School of Information Systems
Singapore Management University
Republic of Singapore
dtnguyen.2011@msis.smu.edu.sg

Hoong Chuin Lau
School of Information Systems
Singapore Management University
Republic of Singapore
hclau@smu.edu.sg

***Abstract*—Unprecedented pace of urbanization and rising income levels have fueled the growth of car ownership in almost all newly formed megacities. Such growth has congested the limited road space and significantly affected the quality of life in these megacities. Convincing residents to give up their cars and use public transport is the most effective way in reducing congestion; however, even with sufficient public transport capacity, the lack of last-mile (from the transport hub to the destination) travel services is the major deterrent for the adoption of public transport. Due to the dynamic nature of such travel demands, fixed-size fleets will not be a cost-effective approach in addressing last-mile demands. Instead, we propose a dynamic, incentive-based mechanism that enables taxi ride-sharing for satisfying last-mile travel demands. On the demand side, travelers would register their last-mile travel demands in real-time, and they are expected to receive ride arrangements before they reach the hub; on the supply side, depending on the real-time demands, proper incentives will be computed and provided to taxi drivers willing to commit to the last-mile service. Multiple travelers will be clustered into groups according to their destinations, and travelers belonging to the same group will be assigned to a taxi, while each of them paying fares considering their destinations and also their orders in reaching destinations. In this paper, we provide mathematical formulations for demand clustering and fare distribution. If the model returns a solution, it is guaranteed to be implementable. For cases where it is not possible to satisfy all demands despite having enough capacity, we propose a two-phase approach that identifies the maximal subset of riders that can be feasibly served. Finally, we use a series of numerical examples to demonstrate the effectiveness of our approach.**

***Keywords*-urban transportation, ride sharing mechanism**

I. Introduction

In the past few decades we have witnessed unprecedented pace of urbanization across the globe. The massive scale of urbanization, rises of income levels, and increasingly affordable cars, have jointly contributed to growing trends of household automobile ownership. And in many newly formed mega-cities, this has created unbearable congestions. To fight these increasingly unmanageable urban congestions, urban planners are quickly expanding public transport network, and are looking for ways to convince people to give up their cars and use public transport instead. For many urban dwellers, one major deterrence in utilizing public transport for their daily commutes is the need for the last-mile (LM) transport, which refers to the travel from the station to the final destination.

A straightforward idea to satisfy the LM demands is to establish a service fleet for each major transport hub. However, due to the fact that the demands for the LM transportation are irregular and distributed (both spatially and temporally), having a fixed-size service fleet is infeasible, for the following intuitive reasons:

1) Demands are highly irregular and uncertain. Therefore, to ensure that the fleet can cope with peaks in demands, the fleet has to run with spare capacity that would be underutilized most of times.
2) To ensure reasonable quality of service, the routes of the fleet have to sufficiently cover most of the service area (the travel time from any point in the area to the closest stop should be within certain minutes) with reasonable service intervals (this constrains longest waiting time). The fleet can operate statically with fixed routes, or it can operate dynamically with routes depending on customers on board; however, in either case, significant slacks have to be introduced in the fleet so as to handle the spatial and temporal demand uncertainties.

Because of the above two issues, operating fixed-size fleets is cost-ineffective for most occasions except for the very limited cases where demands are are consistently high.

A powerful idea in addressing unpredictable travel demands is *sharing*, or *resource pooling*. For example, in many European countries, the bike sharing and car sharing schemes have been suggested as a way to bridge the gaps of public transport. In these instances, resources (bikes and cars) are pooled at fixed locations, and travelers will grab resources when necessary to complete their travels. In this case, resources are pooled and resource utilizations are independent. On the contrary, resources may be independent while the utilizations are pooled. Ride sharing (car-pooling or taxi-pooling) is a typical such case.

In this paper, we propose a formal framework for organizing the last-mile service that is based on non-dedicated fleets (e.g., taxis). In particular, for a single-hub, single-batch scenario with known demands, we specify the conditions

under which all riders and drivers would voluntarily stay with the service.

II. Related Works

The problem of organizing last-mile service has been studied in the literature under various names. The most well-known one is what researchers called the dial-a-ride problem (DARP). DARP is a well-studied hard optimization problem with many variants, and many solution approaches have been proposed in the past (see [1] for a typical exact solution approach for solving DARP; for comprehensive survey, see [2]). Another similar problem is studied in the context of ride sharing, in which passengers and drivers are matched in real time. A simulation study in the city of Atlanta has been recently reported to have good results [3]. The problem of dynamic pickup and delivery is also closely related to our model [4].

Although part of our problem is similar to the DARP and the ride-sharing problem, there are a number of fundamental differences between our model and the models proposed in the literature. First, the last-mile services are mostly organized at transport hubs, with demands coming in batches. By exploiting this features, we can significantly improve the efficiency of the resulting optimization model. Second, we have put major emphases on designing proper incentives for both drivers and riders. As the last-mile service has to be constructed and used voluntarily on both supply and demand sides, the incentive design is thus extremely important to make service sustainable.

III. Organizing the Last-Mile Service

The last-mile (LM) service can be organized under a wide variety of circumstances. In this paper, we assume that there is a single hub, and all demands are with identical departure time from the hub. The destinations of all demands are also assumed to be known and within certain radius from the hub.

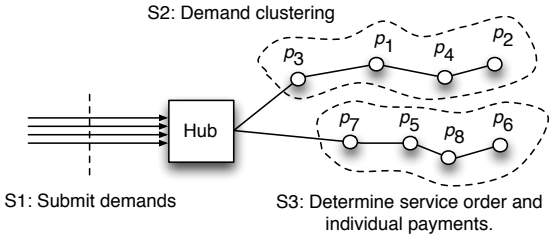


Figure 1. Organizing the last-mile service.

A typical cycle of the LM service can be seen in Figure 1. There are three important steps in organizing the LM service:

1) LM service is organized at a particular major hub where regular train or metro services will be bringing in potential riders at short intervals. For riders who plan to arrive at the hub and utilize the LM service, they have to submit their intents some minutes before their arrivals. It's assumed that all riders will depart immediately for the LM service when they reach the hub, and they will provide the exact coordinates of their destinations.
2) After receiving all destinations at the cut-off time, the central controller should optimally assign all riders to appropriate clusters, where each cluster is to be served by a participating driver.
3) The order of service and the payment to be made by each rider will be decided as we finalize the cluster assignment.

Based on above descriptions, there are two critical problems that need to be repeatedly solved:

1) Demand clustering: In which demands are clustered into groups to be served by different vehicles.
2) Service sequencing and pricing: In which the service order and the associated price for riders in every cluster is determined.

These two problems are closely connected since the planned route and the prices associated with a cluster are highly dependent on the assigned riders. In the following section, we will formally define the clustering model and the service sequencing and pricing model, highlighting how we can explicitly connect these two models.

IV. The Model

The classical DARP is formulated as a mixed integer program (MIP) [1]. Our model is based on the MIP formulation of the classical DARP, but with significant modifications. Our changes are made in order to address the two major differences between the DARP and the LMP: 1) because all riders depart from the same hub all at the same time, we manage to drop the cluster index from all decision variables, and 2) the LM service is based on voluntary participation from both drivers and riders, we thus have to include additional constraints to ensure that the assignment we suggest is dominant choice for all participants. The first change allows us to shrink the solution space roughly by a factor of K (the size of the service fleet), and the second change makes the solution space more constraints, thus our LMP formulation ends up much more compact.

A. Notations and Decision Variables

Let n denotes the number of destinations. Let $G = (N, A)$ be the complete directed graph storing distances between all pairs of destinations (including the hub), where $N = \{0, 1, \ldots, n\}$ represents all drop-off points, and 0 represents the hub node. Let K denote the set of all vehicles. Let Q be the capacity of all vehicles. For arc $(i, j) \in A$, let d_{ij} be the distance required to traverse it.

Define binary variable x_{ij} to be 1 if any vehicle travels on arc (i, j), and 0 otherwise (this is where the cluster index

is dropped, as a result, the number of binary variables is dropped to n^2 from Kn^2). B_i is the travel distance for rider i to reach destination. a_i is the the order of service for rider i in the cluster she is assigned to. p_i is defined to be the price paid by rider i. Finally, let α be the worst-case ratio between real travel distance and direct travel distance; by definition, we can view α as the proxy for the worst-case quality of service (QoS).

B. The Clustering Constraints

The constraints in our LMP can be classified into two major groups. The first group is related to the proper forming of a cluster. The second group is related to the assurance that the obtained assignment will be the dominant choice for all participants.

For a cluster assignment to be valid, the following constraints have to be satisfied:

$$\sum_{i\in N}\sum_{j\in N} x_{ij} = \min\{|N|, |K|\cdot Q\}, \tag{1}$$

$$\sum_{i\in N} x_{ij} \le 1, \forall j \in N, \tag{2}$$

$$\sum_{j\in N} x_{ij} \le 1, \forall i \in N, \tag{3}$$

$$\sum_{j\in N} x_{ij} \le \sum_{h\in N} x_{hi}, \forall i \in N, \tag{4}$$

$$\sum_{j\in N} x_{0j} = |K|, \tag{5}$$

$$a_0 = 0, \tag{6}$$

$$a_j \ge a_i + 1 - M(1 - x_{ij}), \forall i, j \in N, \tag{7}$$

$$a_j \le a_i + 1 + M(1 - x_{ij}), \forall i, j \in N, \tag{8}$$

$$a_i \le Q, \forall i \in N, \tag{9}$$

$$B_j \ge B_i + d_{ij} - M\left(1 - x_{ij}\right), \forall i, j \in N, \tag{10}$$

$$B_j \le B_i + d_{ij} + M\left(1 - x_{ij}\right), \forall i, j \in N, \tag{11}$$

$$B_i \le \alpha \cdot d_{0i}, \forall i, j \in N, \tag{12}$$

$$B_i \ge 0, \forall j \in N, \tag{13}$$

$$x_{ij} \in \{0, 1\}, \forall i, j \in N, \tag{14}$$

$$\alpha \ge 1. \tag{15}$$

We use (1) to ensure that we always serve as many riders as possible. (2) and (3) ensure that there will be at most one vehicle going in to and out of a node (in other words, a rider can only be assigned to one cluster). (4) is for flow conservation: the outgoing flow cannot exceed incoming flow. (5) ensures that there can be exactly $|K|$ departures from the hub (we assume that the fleet size is never larger than the number of requests). The value of a_i (service order for rider i) is characterized by constraints (6) – (9). (6) sets the order of the hub to 0. For (7) and (8), they are equivalent to the non-linear constraint (M is a large constant):

$$a_j = x_{ij}(a_i + 1),$$

which ensures that the service order of j should be one greater than i only if a travel is made from i to j (i.e., $x_{ij} = 1$); the constraints are non-binding if $x_{ij} = 0$. (9) enforces capacity limit on all vehicles. The value of B_i (the real distance traveled by rider i) is characterized by constraints (10) and (11), which are again equivalent to the following non-linear constraint (M is a large constant):

$$B_j = x_{ij}(B_i + d_{ij}),$$

which ensures that j's traveled distance is exactly d_{ij} farther if the vehicle serves i before serving j ($x_{ij} = 1$); the constraints are non-binding if $x_{ij} = 0$. The worst-case QoS, α, is determined by (12). The domains of all related decision variables are specified by (13) – (15).

C. The Rationality Constraint and the Objective Function

As argued earlier, an important property of the LM service is that we need voluntary participations from both the drivers and the riders. It is thus very important to ensure that the suggested assignment and service orders are aligned with all participants' utility functions.

$$\sum_{i\in N} p_i = \delta \sum_{i\in N}\sum_{j\in N} x_{ij} d_{ij}, \tag{16}$$

$$\delta\, d_{0i} \ge p_i + \Delta_i(t_i - s_i), \forall i \in N. \tag{17}$$

The first constraint, (16), ensures that the total price paid by all riders (left-hand side) will be enough to pay all drivers (right-hand side). The δ in the RHS is the paid rate per unit of distance traveled. In our formulation, we assume a simple linear function to transform distance traveled into revenue, but it will be straightforward to incorporate more complicated revenue function. The second constraint, (17), ensures that it's individually rational to participate in the LM service for each and every rider. The LHS is the cost for traveling alone in taxi; the RHS is the actual cost paid plus time penalty resulting from ride sharing. The parameter Δ_i is a rider-specific parameter to convert additional travel distance into monetary penalty.

Finally, the objective function is written such that the total cost paid by all riders is minimized (including both the monetary payments and time penalties):

$$\min \sum_{i\in N} \left(p_i + \Delta_i(B_i - d_{0i})\right) \tag{18}$$

The MIP formulation of the LMP can be defined to have (18) as the objective function, and (1) – (17) as constraints. We will refer to this formulation as **Problem A** for the rest of the paper.

V. Ensuring Implementability

When we solve Problem A and obtain an assignment plan, we will have the following information for each and every rider: 1) the assigned cluster (driver), 2) the order of the service, 3) the arrival time at the destination, and 4) the payment. Naturally, we would want to ensure that the generated plan can be implemented; in other words, all drivers should be content with the income earned (by serving assigned clients), and all riders should be satisfied with the resulting travel times and payments. Defined formally, an implementable plan must meet the following criterion:

- **Budget balance**. Payments from riders should provide drivers with sufficient monetary incentives to stay with the LM service. This holds by construction due to (16).
- **Individual rationality**. All riders should prefer using the LM service than traveling alone. This holds by construction due to (17).

From the above criterion, Problem A, if feasible, generates assignment plan that is guaranteed to be implementable.

The demand pattern illustrated in Figure 2 is an example where no implementable sharing plan exists. No matter how small the Δ_i value is (unit penalty for extra travel time), riders still cannot feasibly travel in group, since payments affordable for riders are not enough to pay drivers collectively.

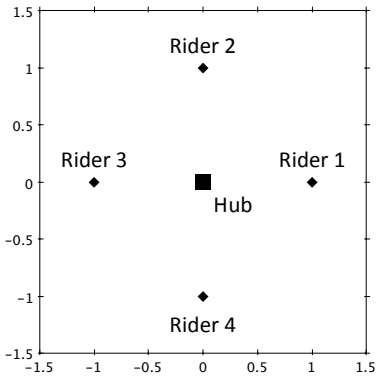


Figure 2. An example with four riders, but can only be feasibly served with four drivers.

VI. Handling Infeasible Set of Riders

As noted previously, Problem A might contain no feasible solution even when we have enough capacity in the fleet (an example of this is already illustrated in Figure 2). However, not able to feasibly solve Problem A doesn't mean that a feasible LM service cannot be formed; a feasible LM assignment can still be formed if a proper subset of the riders can be dropped. Again, by using the same example in Figure 2 (let's assume one driver is available), we can see that a feasible assignment can still be formed if we drop riders 2, 3, and 4 and only serve rider 1 (the solution, which can be easily obtained in this example, can only be found by re-solving the Problem A in general).

To formalize the idea, we define a two-phase procedure to select a proper set of riders to be eventually served. The first phase serves the purpose of filtering riders, and it can be achieved by solving a variant of the Problem A, which we call it the Problem B. Problem B is only slightly different from Problem. The set of constraints stays mostly the same, except for Equation (1), which is modified to:

$$\sum_{i\in N}\sum_{j\in N} x_{ij} \leq \min\{|N|, |K|\cdot Q\}. \tag{19}$$

The above modification allows us to serve below fleet capacity or drop some riders. To ensure that Problem B still tries to serve as many riders as possible, we also modify the objective function to be:

$$\max \sum_{i,j} x_{ij} + \frac{1}{\sum_j d_{0j}}\Big(\sum_{i,j\in N} x_{ij} d_{0j}\Big). \tag{20}$$

The first part of the objective function is simply the number of riders served. The second part of the objective function is the average normalized direct distance of all served riders. By combining these two components, our first priority is serve as many riders as possible; when there are more than one assignments that allow us to serve the same number of riders, we would prefer serving riders with longer travel requests.

With Problem B, the two-phase procedure can formally be implemented as follows:

1) (*Phase I*) Solve Problem A, if the problem is feasible, stop; otherwise, move the Step 2.
2) (*Phase IIA*) Solve Problem B, obtain the subset of riders that are served in Problem B (to discover riders that are chosen to be served in Problem B, simply find all j such that $x_{ij} = 1, \forall\, i$).
3) (*Phase IIB*) Configure Problem A to include only riders that are served in Problem B. Re-solve Problem A to obtain the assignment tuple.

VII. Experimental Results

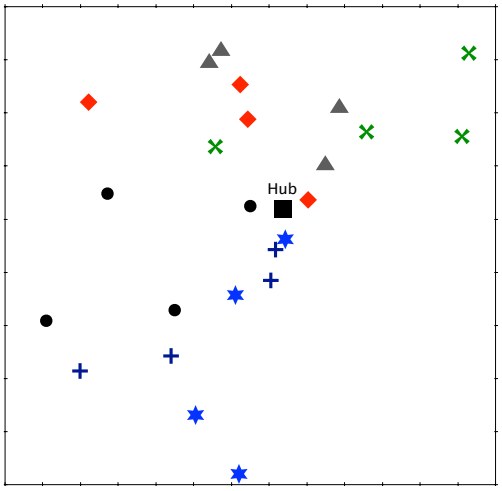
Table I
Summary of the LM planning results for different problem sizes (all Δ_i are set to 1).

Riders	Drivers	Time / Time using DARP Model	α	Total Distance	Extra Travel
8	4	0.37s / 43s	1.13	242.3	13.9
20	5	24.06s / 3m14s	1.79	467.0	188.6
24	6	1m24s / no result after 3h	1.75	522.6	210.4
32	8	17m22 / no result after 1d	1.55	628.2	189.2

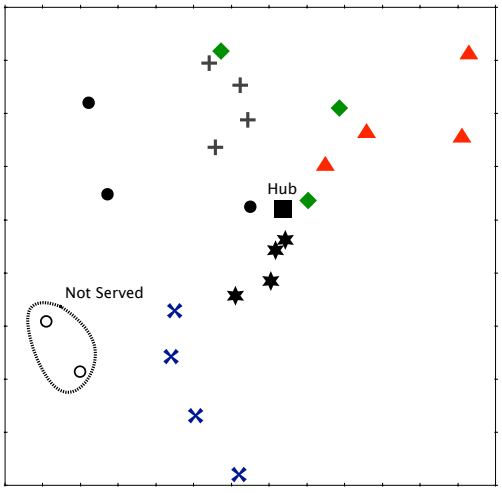
Table I summarizes the performance statistics we obtain under different problem sizes. The first thing to note is

the significantly improved solution speed. For the largest instance, our model (Problem A) returns solution within 17.5 minutes, while the classical DARP MIP model runs over one day without terminating. For the largest instance, our formulation is at least two orders of magnitude faster than the classical DARP MIP model. Also note that all results are obtained assuming that $\Delta_i = 1$ for all i. By changing Δ_i, we may obtain different results; most significantly, the additional travel should reduce as a result.

Another interesting result is the dropping of riders using the two-phase procedure. Use the case with 24 riders as example: by setting $\Delta_i \geq 3$, we will begin to see riders being dropped. The clustering results with $\Delta_i = 1$ and $\Delta_i = 3$ are illustrated in Figure 3. We can see significant difference in additional travel distance.



(a) The case with $\Delta_i = 1$.



(b) The case with $\Delta_i = 3$.

Figure 3. The clustering result with different Δ_i values.

VIII. Conclusions and Future Work

In this paper, we show that a LM service fleet can be organized dynamically by tapping into spare capacity of a taxi fleet. With the assumptions that demands are known *a priori* and come in batches, we demonstrate that a LM service can be organized for a single batch of demand by solving a MIP model. The most important feature of our model is the embedded implementability constraints, which guarantees that the obtained solution will always be implementable. For cases where feasible solution does not exist, we devise a two-phase procedure where a promising subset of riders can be chosen to be served.

Through a series of numerical examples, we show that our approach can obtain solution at least two orders of magnitude faster when compared against classical DARP MIP model. We also demonstrate how we can control the performance of overall fleet by adjusting Δ_i. The effectiveness of our two-phase approach is also demonstrated.

There are two major areas that we would like to further develop. First is the handling of more complicated demand scenarios. Second is the analyses on behavioral and societal impacts. For the first area, we are interested in addressing multiple batches of demands, which can be assumed to be known (pre-registered using the same technology) or partially known (uncertain). Also, we would like to address the issue of last-minute changes: e.g., handling re-clustering to handle no-show or walk-in riders. Finally, we would like to include not just LM, but also the first-mile (FM) service, which goes in the reverse direction from an arbitrary origin to a hub. For the second area, we will assess the potential impacts of the LM service on other types of transport service, e.g., the feeder bus or light-rail system in the neighboring area. On behavioral issues, we would like to address rider's preference in demand clustering, e.g., riders might prefer not sharing rides with more than certain number of people, or they may prefer to share the rides only with certain gender or age groups.

References

[1] J.-F. Cordeau, "A branch-and-cut algorithm for the dial-a-ride problem," *Operations Research*, vol. 54, no. 3, pp. 573–586, 2006.

[2] J.-F. Cordeau and G. Laporte, "The dial-a-ride problem (DARP): Variants, modeling issues and algorithms," *4OR: A Quarterly Journal of Operations Research*, vol. 1, pp. 89–101, 2003. [Online]. Available: http://dx.doi.org/10.1007/s10288-002-0009-8

[3] N. A. Agatz, A. L. Erera, M. W. Savelsbergh, and X. Wang, "Dynamic ride-sharing: A simulation study in metro Atlanta," *Transportation Research Part B: Methodological*, vol. in press, 2011.

[4] R. Baldacci, E. Bartolini, and A. Mingozzi, "An exact algorithm for the pickup and delivery problem with time windows," *Operations Research*, vol. 59, no. 2, pp. 414–426, 2011.