A Graph-Based Approach to Measuring the Efficiency of an Urban Taxi Service System

Xianyuan Zhan, Xinwu Qian, and Satish V. Ukkusuri

***Abstract*—Taxi service systems in big cities are immensely complex due to the interaction and self-organization between taxi drivers and passengers. An inefficient taxi service system leads to more empty trips for drivers and longer waiting time for passengers and introduces unnecessary congestion on the road network. In this paper, we investigate the efficiency level of the taxi service system using real-world large-scale taxi trip data. By assuming a hypothetical system-wide recommendation system, two approaches are proposed to find the theoretical optimal strategies that minimize the cost of empty trips and the number of taxis required to satisfy all the observed trips. The optimization problems are transformed into equivalent graph problems and solved using polynomial time algorithms. The taxi trip data in New York City are used to quantitatively examine the gap between the current system performance and the theoretically optimal system. The numerical results indicate that, if system-wide information between taxi drivers and passengers was shared, it is possible to reduce 60%–90% of the total empty trip cost depending on different objectives, and one-third of all taxis required to serve all observed trips. The existence of destructive competition among taxi drivers is also uncovered in the actual taxi service system. The huge performance gap suggests an urgent need for a system reconsideration in designing taxi recommendation systems.**

***Index Terms*—Taxi service system, large-scale trip data, efficiency, graph theory, minimum weight perfect bipartite matching, minimum path cover.**

I. Introduction

THE rapid deployment of various mobile sensors have steadily increased the quantity of data available in various systems, documenting the movements of people, bits, and ideas at a rate that was unimaginable before. The ability to collect, interact and analyze massive streaming data has unambiguously transformed our understanding of many extremely complex systems that were impossible to be completely modeled before. Instead of a static snapshot of the underlying phenomenon, technology equipped networked agents provide disaggregate data of their location and state allowing the characterization of dynamics of the system.

In urban systems, taxis serve as an indispensable mode of transportation for point-to-point travel. Typical urban areas picking up and dropping off. By the end of 2013, there were 13 437 yellow medallion taxicabs in New York City (NYC) that transport more than 236 million passengers [1]. In Hong Kong, there are about 15 000 taxicabs and more than a million passengers served everyday [2]. While up to 60% of the daily traffic flow in Hong Kong are generated by taxis, there are significant number of empty trips [3]. The excessive empty trips increase the waiting time for passengers and the avoidable operating cost for drivers, which eventually lead to a series of negative externalities, such as urban congestion and emissions [4]. Maintaining a high system operational level and increasing utilization of the taxi service system are of key concerns faced by many cities.

Previous approaches address the inefficiency issue of the taxi service system at the aggregate level (total demand and supply), where economic relationships are considered to determine the optimal fare setting and fleet size [3], [5]–[8]. All these studies provided a preliminary direction to enhance the system performance by introducing entry and fare controls to the taxi market. The limitation, however, comes from the unrealistic assumption of the taxi driver and passenger's behavior and the inability to fully characterize the taxi service system using overly abstracted mathematical models. The taxi service system is an immensely complex self-organized system: drivers are self-adaptive based on their own knowledge of the traffic, and the passenger demand is both spatially and temporally varying. For example, drivers are likely to roam near residential areas during morning peak and wait at the concert when approaching the end of a play. The inefficiency of the system arises, even when the market is properly regulated, due to the lack of perfect system-wide information shared between taxi drivers and passengers.

Taxis in urban areas such as NYC are equipped with GPS devices, providing second by second location information. The unprecedented amount taxi trip data generated from GPS equipped taxis allow researchers to directly observe and analyze the system performance and recommend various taxi related services. Further, taxis serve as probes in the road network and provide massive amounts of streaming data, and the analysis of which provides important performance metrics to understand causal factors for taxi ridership [9], dynamics of taxi demand [10], [11] and to estimate real-time speed [12], [13]. Consequently, new approaches emerge, such as dispatching system, taxi recommendation and ridesharing applications, which utilize the real-time trip information or historical taxi trip data to provide various user services [14]–[21]. By clustering historical pick-up locations based on temporal and spatial characteristics, guidances are provided to help drivers reduce the number of

Manuscript received April 19, 2015; revised July 23, 2015, October 5, 2015, and December 13, 2015; accepted January 11, 2016. The Associate Editor for this paper was D. Srinivasan.

The authors are with the School of Civil Engineering, Purdue University, West Lafayette, IN 47906 USA (e-mail: sukkusur@purdue.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2016.2521862

empty trips. The cruising distance before finding a passenger can be reduced by learning from experienced drivers and making a sequence of recommendations [16]. Yamamoto *et al.* [17] proposed a fuzzy clustering and adaptive routing algorithm to dispatch vacant taxis to places where passengers are more likely to be found. In addition to aid taxi drivers, Yuan *et al.* [18] also incorporated passengers' mobility patterns and taxi driver's pick up/drop off behaviors to provide recommendations for passengers to reduce disequilibrium between the demand and supply. Apart from improving the taxi services through dispatching passenger and vacant taxis, or providing guidance to taxi drivers, ridesharing services is another way to reduce congestion and energy consumption. Several works on ridesharing systems [19]–[21] have been developed with the consideration of time, capacity and monetary constraints. Despite the technological advances in ridesharing services, non-technical problems remain, such as legitimate issues as well as security issues [14]. However, a fundamental scientific question to be answered is how efficient is the current taxi service system, and how to quantitatively evaluate the performance level of the current system. Moreover, if the optimal performance of the taxi service system can be quantitatively measured, how far is the performance gap between the current system and the theoretical optimum? The inefficiency arises largely due to the lack of globally shared information among taxi drivers and passengers. In addition, the local greedy choices that drivers make to pick up passengers could also result in system wide inefficiencies. While some drivers or passengers may benefit from the aforementioned taxi dispatching and recommendation schemes, whether the efficiency of the entire system is improved remains questionable. Intervention strategies without considering system-level efficiency may lead to downgraded system performance.

In this article, we evaluate the efficiency level of the taxi service system and quantitatively measure the theoretical optimal performance of the system. We analyze the entire taxi service system using equivalent graph representations. The dataset used in this study contains over 500 000 daily trip records from the real world, which is an ideal source to inspect the interaction between taxi drivers and passengers. The system efficiency is explored following two schemes: (1) the optimal matching and (2) the trip integration. Given the sets of available taxis and passengers within a time interval, the optimal matching provides the best matching strategy between vacant taxis and possible passengers. For solvability, this is transformed to a minimum weight perfect bipartite matching problem. On the other hand, given the information for a set of taxi trips, the trip integration finds the optimal integration strategy of sequences of trips for the taxis in the system, which is shown to be equivalent to a minimum path cover problem. Our results suggest that the taxi service system of New York City has a significant performance gap compared with the theoretical optimal system due to the lack of system wide information shared between drivers and passengers. While there are other factors that may prevent from attaining the theoretical optimum, this provides a clear benchmark of the potential efficiency gains that can be realized by introducing new technology that allowing sharing system-wide information in the taxi system. Furthermore, if perfect information is provided, only two thirds of all taxis are sufficient to serve all observed trips and under certain scenarios, up to 90% of total empty trip cost may be reduced.

The rest of the paper is structured as follows. The next section describes the data used for this paper; Section III presents the methodological approach developed to find the theoretically optimal strategy; Section IV shows the experiment results and the concluding remarks are given in the final section.

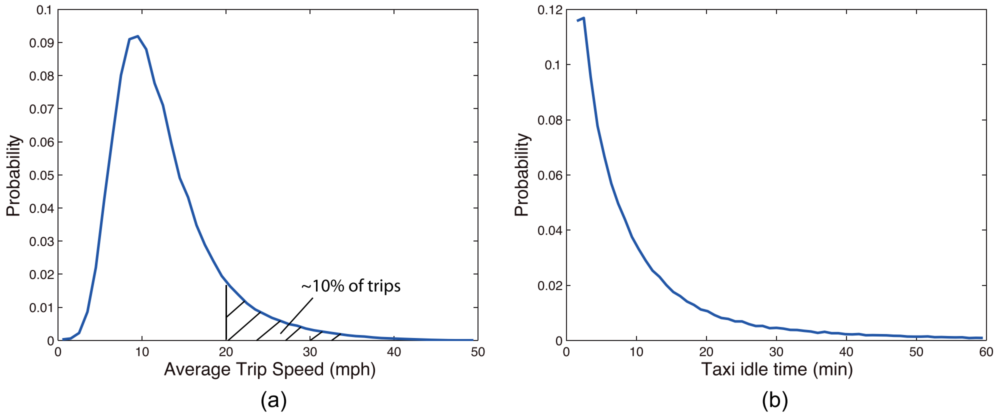
II. Data

The data used in this research were collected by the New York City Taxi and Limousine Commission (NYCTLC) on the trip by trip basis. The data contain the taxi medallion ID, driver initial, shift number, the timestamp and GPS coordinates of trip origin and destination, trip duration, travel distance and fare etc. Around 400 000 to 500 000 daily trips were recorded during the data collection period from December 2008 to January 2010. Observing the stable trip pattern during weekdays and weekends [22], one-week data from October 5th, 2009 to October 11th, 2009 were extracted for further analysis, in which no major social events were reported. We focus on investigating the efficiency of the taxi service system using three days' data from this week (Wednesday 2009-10-07, Friday 2009-10-09 and Saturday 2009-10-10).

A. Data Pre-Processing and Macroscopic Characteristics

There are more than 13 000 medallion cabs in NYC (NYCTLC, 2012), but only 818 unique medallion IDs are found in the data, which is far from the right number. Analysis on the data have shown extensive reuse of the same medallion ID numbers for different taxis exists in the dataset. In the data, a specific medallion ID is associated with multiple driver initials and shift numbers. It has been observed that multiple trip records with the same medallion ID but with different shift numbers and driver initials can occur at the same time. Consequently, the medallion ID is not a reliable identifier for taxis in the dataset. To be rigorous, this study used the concatenated string of taxi medallion ID, driver initial and shift number as the unique taxi identifier (referred as taxi identification number). For the weekday data, we observe a total of 489 234 trips within a day, and 32 368 (6.6% of the total number of trips) distinct taxis identification numbers are recovered. For the weekend data (Saturday), 33 999 taxis are recognized from the 524 792 trips. Considering that there are usually two to three daily shifts per taxi [1], the amount of observed unique taxi identification numbers suggests a range of 11 000 to 17 000 medallion cabs in NYC, which agrees with the reported number of 13 000 by NYCTLC [1].

All trips under a same taxi identification number were retrieved and sorted based on trip starting time. The empty trip information was obtained by comparing the consecutive drop-off and pick-up locations and timestamps. For each taxi, since it is hardly possible to gain information prior to its first trip (referred to as starting trip in following content), only the drop-off location of the first trip is utilized and served as the initial position of the taxi. Some of the related macroscopic statistics of the taxi trip data, including the distribution of average trip speed and taxi idle time are presented in Fig. 1.



(a)

(b)

Fig. 1. Related macroscopic statistics of the taxi trip data. (a) Distribution of average trip speed. (b) Distribution of taxi idle time.

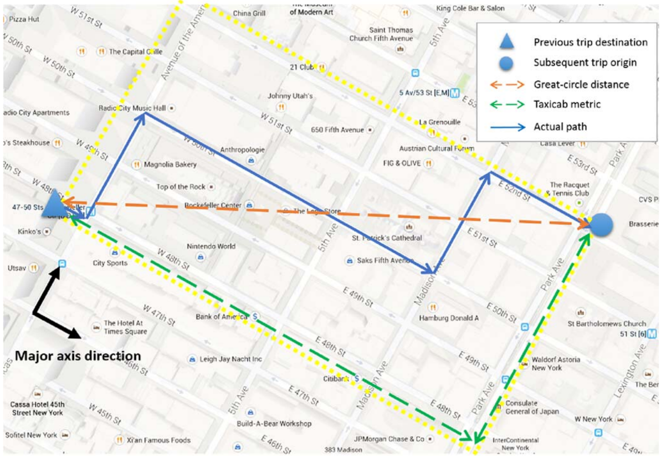


Fig. 2. Illustration of taxicab metric.

B. Distance Measure

Due to the lack of detailed trajectory information in the data, the actual vacant travel distance between successive trips is unknown. One may approximate the empty trip distances as the great-circle distance between successive trip destination and origin computed using haversine formula. However, this method ignores the actual path taken of a taxi and hence underestimate the real empty trip distance. Instead, in this study, we approximate the empty trip distance using taxicab metric (also known as city block distance, Manhattan distance), which achieves higher level of accuracy by fully exploit the special grid-like road network structure in NYC. Fig. 2 illustrates computation of taxicab metric. A rectangular boundary is first constructed using the starting and ending locations under the orientations of the two major axis (obtained by analyzing the orientation of the entire road network data of NYC). The great-circle distances of the two edges of the rectangle boundary are then computed using haversine formula, the sum of which gives the taxicab metric of the empty trip. Note that under a grid structured road network, as long as the trajectory of the empty trip falls within the defined rectangular boundary, its distance will always be the same as taxicab metric.

III. Methodology

This section presents the details of the two approaches used to evaluate the efficiency of the taxi service system, namely the optimal matching and the trip integration. Consider the following two aspects of the efficiency in the taxi service system:

- If the information of trip starting time and location for each passenger, and the current location of each available taxi is known within a time interval, how to match the given set of vacant taxis and the passengers so that the total time/distance or revenue loss are minimized?
- If a given set of taxi trips are known beforehand, how to combine a sequence of trips served by individual taxis to achieve the minimum utilization of the number of taxis and lower the total trip cost?

The first research question focuses on the issue that taxi drivers have to spend too much time or travel extra miles than actually needed to find the next passenger. The second research question is particularly meaningful if there is a high demand for taxis while the supply is very limited, such as during peak hours. Apparently, the system efficiency will be improved by addressing either of the two problems. The optimal matching provides an optimal solution for the first scenario while the second one is addressed by the trip integration.

A. Assumptions

There are four major assumptions related to the methodology:

1) There exists a hypothetical system-wide recommendation mechanism (e.g. an Uber-like taxi hailing APP but adopted by all users in the system) for conducting the optimal matching and trip integration for every time interval T with a length of ΔT, within which the information of both taxis and passenger trips are revealed at the beginning of the time interval.
2) The starting trips are only considered as a starting point and are not involved in the matching cost computation.
3) All trip observations must be served at the exact time and location as recorded in the data.

4) The travel distance between consecutive trips (empty trip distance) is computed in taxicab metric rather than the actually travel distance, due to the lack of detailed trajectory information and the large size of the data.

In this study, the reason for discretizing time into intervals is to mimic the operation behavior of many real taxi recommendation and dispatching systems. Such systems typically discretize time into a set of "batching windows," where the information from both available taxis and passengers is collected inside each batching window and then used to make trip recommendations. The optimal matching requires a short ΔT, while a longer ΔT is expected to have enough trips for the integration.

B. Notation

The notations used in the mathematical formulations are given as follows:

T Index of the current time interval, $T \in \{1, 2, 3 \ldots, \mathbb{T}\}$.

ΔT Length of the time interval.

A^T The set of all available taxis at time interval T. All taxis finish serving a trip within T are seen as available, except the observed last trip of a taxi. Each available taxi in A^T is represented as a tuple: (i, p_i^a, t_i^a), corresponding to the taxi identification number, location and timestamp when available (end point and timestamp of last trip), the superscript a refers to "available taxi."

B^T The set of trips to be served in interval T. Each trip in B^T is represented as a tuple: $(j, p_j^o, t_j^o, p_j^d, t_j^d)$, corresponding to the trip ID, location and timestamp of trip origin, location and timestamp of trip destination respectively. The superscript o and d refer to the origin and destination respectively.

R^T The set of unmatched taxi in time interval T. The cardinality of R^T is $|R^T| = |A^T| - |B^T|$.

M A sufficient large number.

z_{ij}^T A possible matching between taxi $i \in A^T$ and trip $j \in B^T$.

d_{ij}^T The Manhattan distance of the matching between p_i^a and p_j^o.

α, β Cost coefficients for the time and distance of empty trips.

w_{ij}^T Cost for matching taxi $i \in A^T$ and trip $j \in B^T$.

$G(V, E)$ Graph notation with the set of vertices $V = \{v_k\}$ and the set of edges $E = \{e_{\text{ij}}\}$.

c_{ij}^T Capacity of edge $e_{\text{ij}} \in E$.

$v_{\max}$ Maximum travel speed for matching. Set as 20 mile/hour in actual implementation according to Fig. 1(a). The value covers 90% of all trips observed without being too conservative.

C. Optimal Matching

Given the time interval ΔT (assumed to be small enough such that no more than one complete trip is finished during the interval), the objective of the optimal matching is to find the optimal matching strategy between each pair of taxi driver and passenger, so that the total matching cost is minimized. The matching cost can be measured as the taxi idle time, empty trip distance or revenue loss, which is the weighted combination of the previous two. It can be interpreted as a proxy of the empty trip cost spent by a taxi driver in finding the next passenger. We will use the term "matching cost" and "empty trip cost" interchangeably in the following discussion. The optimal matching problem for each interval $T = 1, 2, \ldots, \mathbb{T}$ can be formulated as the following integer linear program (ILP):

$$\text{Min} \sum_{\text{ij}} w_{\text{ij}}^T z_{\text{ij}}^T, \quad i \in A^T,\ j \in B^T \bigcup R^T \tag{1}$$

$$\text{s.t.} \sum_{i} z_{\text{ij}}^T = 1, \quad i \in A^T \tag{2}$$

$$\sum_{j} z_{\text{ij}}^T = 1, \quad j \in B^T \bigcup R^T \tag{3}$$

$$w_{\text{ij}}^T = \begin{cases} \alpha\left(t_j^o - t_i^a\right) + \beta d_{\text{ij}}^T, & \text{if } j \in B^T \\ M, & \text{o.w.} \end{cases} \tag{4}$$

$$z_{\text{ij}}^T \in \{0, 1\}. \tag{5}$$

Equations (2) and (3) ensure the one to one mapping (perfect matching) between taxis and trips. Equation (4) defines the matching cost. Equation (5) restricts z_{ij}^T to be a binary variable. $z_{\text{ij}}^T = 1$ if a matching exist, and 0 otherwise. A matching is *valid* (matching cost w_{ij}^T is finite) if and only if the following conditions are satisfied:

$$t_j^o \geq t_i^a \tag{6}$$

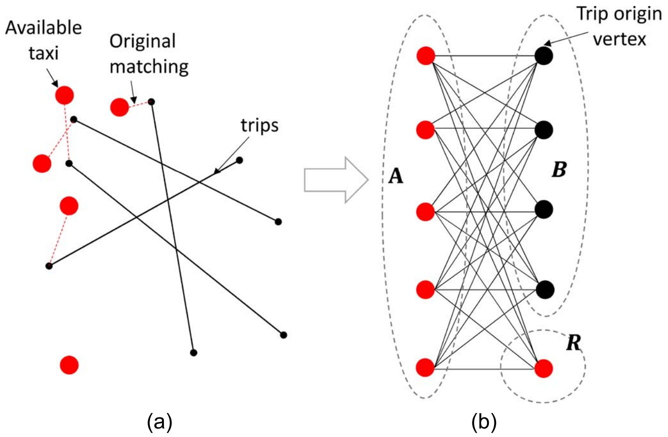
$$d_{\text{ij}}^T \leq \left(t_j^o - t_i^a\right) v_{\max}. \tag{7}$$

Equation (6) states that the taxi available time t_i^a should be no later than the trip starting time t_j^o. Equation (7) sets the maximum possible distance in taxicab metric between available taxi location and trip starting location for a valid matching.

If we abstract the set of taxis and the set of trips as two sets of vertices, and the set of valid matching as the set of edges, the ILP problem can be represented as a bipartite graph as illustrated in Fig. 3(a). Red vertices are the available taxis and black vertices represent the trips to be served. Moreover, there is a cost w_{ij}^T associated with each pair of matched taxi and passenger trip, which is considered as the weight of the corresponding edge. We can show the ILP is equivalent to the minimum weight perfect bipartite matching problem illustrated in Fig. 3(b).

Definition 1: Given a bipartite graph $G = (V, E)$ with bipartition S and T $(V = S \bigcup T)$ and weights w_{ij} for all edges $e_{\text{ij}} \in E$ $(E = S \times T)$, a **bipartite matching** is a subset of edges $M \subseteq E$, such that for all vertices $v \in S \bigcup T$, at most one edge of M is incident on v. A matching is **perfect** if no vertex is exposed. The **minimum weight perfect bipartite matching** is to find a perfect matching of minimum cost.

From above definition, it can be easily verified that the ILP defined by Equations (1)–(5) is exactly equivalent to the



(a)

(b)

Fig. 3. Graphical representation of trip matching.

minimum weight perfect bipartite matching problem. Note that the optimal matching problem has a nonempty feasible solution set, as at least one matching (the one in real world) exists for the problem. Hence the given problem is always solvable, and there is at least one available taxi for each trip (indicates $|A| \geq |B|$, and $|R^T| = |A^T| - |B^T|$), thus the perfect matching is always well defined. Note that when applied to real world cases, the perfect matching may not always exists, as $|A| < |B|$ may occur. However the problem can still be solved as a **minimum weight maximum bipartite matching** problem. Essentially, this is a reduced problem of optimal matching proposed in the paper, as we do not need the set of unused taxis and the matching needs not to be "perfect," which is easier to solve. The Hungarian method [23] is known to solve the class of minimum weight bipartite matching problem, which has a complexity of $O(|V|^3)$ [24]. For more information about minimum weight bipartite matching and the Hungarian method, please refer to [25].

By solving the optimal matching problem and obtain the optimal solution $z_{\text{ij}}^{T^*}$, the optimal matching strategy can be retrieved and the set of available taxis A^T is updated for the next interval $T+1$ as follows:

1) If $j \in B^T$, then taxi i is matched to trip j. If trip j is not the last trip of a taxi, set $t_{i'}^a = t_j^d$ and $p_{i'}^a = p_j^d$, where $i' \in A^{\tilde{T}}$ and $\tilde{T}$ is the time period t_j^d belongs. The optimized empty trip idle time and distance are computed as $t_k^* = t_j^o - t_i^a$ and $d_k^* = d_{\text{ij}}$.
2) If $z_{\text{ij}}^{T^*} = 1, j \in R^T$, then taxi i is not matched to any trip and is kept for next time interval. Set $t_{i'}^a = t_i^a$ and $p_{i'}^a = p_i^a$, where $i' \in A^{T+1}$.

For each trip $k = 1, 2, 3, \ldots, n^T$, in which n^T is the total number of trips for time interval T, the total matching cost is calculated as:

$$\sum_{T=1}^{\mathbb{T}} \sum_{k=1}^{n^T} \alpha t_k^* + \beta d_k^*. \tag{8}$$

Specifically, if:

1) $\alpha = 1, \beta = 0$: the objective is to minimize the total taxi idle time;
2) $\alpha = 0, \beta = 1$: the objective is to minimize the total empty trip distance;
3) $\alpha = \alpha_0, \beta = \beta_0$: the objective is to minimize the total revenue loss from empty trips, where α_0, β_0 are the cost coefficients of time and distance components.

TABLE I
LINEAR REGRESSION MODEL FOR TAXI FARE-TIME-DISTANCE RELATIONSHIP

	Coefficient	Standard Deviation	P-value
constant	2.143	0.0016	0.000
α_0 (min)	0.275	0.0002	0.000
β_0 (mile)	1.563	0.0006	0.000
Number of observations		415561	
R^2		0.99	

We use a simple linear regression model for taxi fare with the dependent variable of the travel time and distance to measure the empty trip cost. The linear regression model was built using 415 561 taxi trip records and presented in a previous study of the author [12]. The linear model fits the data well, with highly significant parameters and a $R^2 = 0.99$. The model is presented as follows and the results are presented in Table I:

$$\text{fare} = \alpha_0 \cdot \text{time} + \beta_0 \cdot \text{distance}. \tag{9}$$

D. Trip Integration

Given a time interval ΔT and a set of observed trips, the objective of the trip integration is to find an optimal trip combination (integration) strategy that: (1) results in the minimum number of taxis required to satisfy all the trips (*unweighted trip integration*); and (2) results in minimum total matching cost while achieving minimum possible number of taxis satisfying all the trips (*weighted trip integration*). The notion of trip integration is different from the usual trip merging or combination in ridesharing problems [26]. The ridesharing problem typically focuses on combining multiple taxi trips on a similar path using a shared taxi. However, we focus on integrating successive trips for each individual taxi and the taxi sharing behavior is not considered, as it is not revealed in the real-world taxi trip data. The trip integration can be especially beneficial in cases such as peak hours, when the available taxis are not sufficient to address the overflow of passenger trips. By introducing the trip integration, the resources for taxis will be fully utilized and the system output is maximized.

For trip integration, a longer ΔT (e.g., 10 min) is needed compared with the optimal matching. Since small ΔT will result in fewer trips for integration, which may lead to limited improvement to the system. We assume at the beginning of each time interval, all passengers provide their trip information and the trip travel times are known (in this paper) or can be accurately estimated (in real world implementations). Two rules are introduced to verify if the two trips $(i, p_i^o, t_i^o, p_i^d, t_i^d)$ and $(j, p_j^o, t_j^o, p_j^d, t_j^d)$ are possible to be combined (integrated), and we call such trips *combinable trips* if:

1) $0 \leq t_j^o - t_i^d \leq \Delta D$, where ΔD is the maximum delay allowed;
2) $d_{\text{ij}}^T \leq (t_j^o - t_i^d) v_{\max}$, that the taxi is possible to reach the passenger given the observed distance and time.

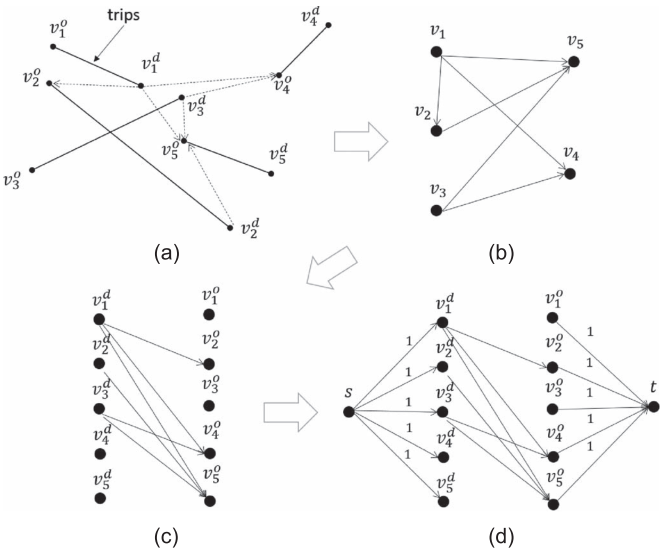


Fig. 4. Illustration of Unweighted Trip Integration. (a) $G(V, E)$. (b) $G'(V', E')$. (c) $G''(V'', E'')$. (d) $G'''(V''', E''')$.

In order to formulate and solve the described optimization problems, we transform the original problems into corresponding graph problems. Particularly, we will consider two cases: (1) unweighted trip integration, which finds the minimum number of taxis required to satisfy all the trips; and (2) weighted trip integration, which finds the minimum total matching cost while achieving minimum possible cardinality set of taxis to satisfy all the trips.

1) Unweighted Trip Integration: We first transform the unweighted version of the problem into a graph representation. Let the abstracted vertex $v_i^o = (p_i^o, t_i^o)$ and $v_i^d = (p_i^d, t_i^d)$ as the origin and destination location and time tuple of trip i. For each trip, there is an edge connecting v_i^o and v_i^d. Furthermore, we add an directed edge between v_i^d and v_j^o if trip i and trip j are combinable (represented as dash line in $G(V, E)$ of Fig. 4(a)). Hence the original unweighted trip integration problem is to find a set of connecting edges between trips that form a set of disjoint paths covering all the trips with the minimum cardinality. If we abstract each trip as a single vertex, let $v_i = (v_i^o, v_i^d)$, and only consider the directed edges that connects combinable trips, then we obtain the directed graph $G'(V', E')$ shown in Fig. 4(b). It can be shown that the unweighted trip integration problem is equivalent to find a *minimum path cover* on G'.

Definition 2: Given a directed graph $G = (V, E)$, the **minimum path cover** is to find the minimum number of paths such that every vertex $v \in V$ belongs to exactly one path. Zero length path is allowed, which is a single vertex.

The equivalency between unweighted trip integration and minimum path cover is straightforward. Since a path is constructed only when every consecutive trip vertices belongs to it are combinable trips. The minimum path cover finds a set of paths that ensure every vertex belongs to a disjoint path, thus all the trips are served, and each trip is served by exactly one taxi. The cardinality is minimal, thus we find the minimum number of taxis that serve all the required trips. The paths found by the minimum path cover will be the optimal integrated trips.

Although the minimum path cover problem is NP-hard in general (as a path cover has cardinality 1 if and only if the directed graph has a Hamiltonian path, which is a NP-complete problem), it is solvable in polynomial time on directed and acyclic graphs. For our problem, since the directed edges connects combinable trip vertices, and by the definition of combinable trips, along any existing path, the trip origin timestamp t_i^o will always be increasing. Consequently, the equivalent directed graph G' will never become cyclic. To solve the minimum path cover problem defined on graph G', we create a equivalent bipartite graph $G''(V'', E'')$ as shown in Fig. 4(c). To perform the transformation, we partition all vertices v_i^d and v_i^o into two sets. The edges that connect the combinable trips between v_i^d and v_j^o are also included. It can be shown that by solving a maximum bipartite matching on G'', we solve the minimum path cover problem. To see this, we introduce following definition and proposition:

Definition 3: A **maximum matching** is a matching M on $G(V, E)$ such that every other matching M' satisfies $|M'| \leq |M|$.

Proposition 1: The directed bipartite graph $G''(V'', E'')$ has a matching of size $n - k$ ($n = |V'|$) if and only if there are k directed paths covering all the vertices in G'.

Proof: Assume the directed acyclic graph $G'(V', E')$ has a path cover P of size k, and let $P_1, P_2, \ldots, P_k$ be the k paths. Hence in the transformed bipartite graph G'', for each path P_i, there will be $|P_i| - 1$ matching edges used. Then there is a matching of size $\sum_i^k (|P_i| - 1) = \sum_i^k |P_i| - k = n - k$.

On the other hand, if G'' has a matching of size $n - k$, then it will form m disjoint paths $P_1, P_2, \ldots, P_m$ and l isolated vertices (both v_i^o and v_d^i are not connected to any other vertices) in G'. Hence these disjoint paths and isolated vertices will form a path cover P on G'. Since $\sum_i^m |P_m| + l = n$ and $\sum_i^m (|P_i| - 1) = \sum_i^m |P_i| - m = n - k$, thus $m + l = k$, and P is a path cover of size k. ■

A well-known solution approach for maximum bipartite matching is to use the max-flow algorithm with a simple graph transformation [27], which is shown in Fig. 4(d). We add dummy source and sink nodes s, t that connect to all v_i^d and v_i^o respectively, and set capacity of all edges to be 1. Solving the max-flow problem on G''' between s and t will find the maximum bipartite matching on G''. According to Proposition 1, finding the maximum matching on G'' will lead to the minimum size of path cover on G' (maximize $n - k$ is equivalent to minimize k). Hence the unweighted trip integration problem can be efficiently solved using polynomial time max-flow algorithm, specifically, if the Edmonds and Karp algorithm [27] is used, the computation complexity is $O(|E''|^2)$.

2) Weighted Trip Integration: If we consider each edge $e'_{\mathrm{ij}} \in E'$ in Fig. 4(b) to be associated with the weight w_{ij}^T, then the problem of finding the minimum number of taxis can be extended to the problem of finding the minimum total matching cost using the least number of taxis. The equivalent graph of this problem is similar to the unweighted case with the adding of weights on edges, which is presented in Fig. 5(a). The objective is to minimize the total weight on the set of disjoint paths that cover all abstracted trip vertices, hence is equivalent to a minimum weight minimum path cover problem.

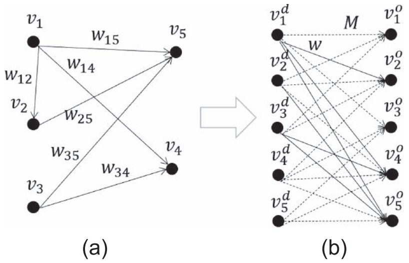


Fig. 5. Illustration of Weighted Trip Integration. (a) $G'_w(V'_w, E'_w)$. (b) $G''_w(V''_w, E''_w)$.

Definition 4: Given a directed graph $G = (V, E)$, the **minimum weight minimum path cover** is a minimum path cover P that minimize the sum of the weights of the edges of paths of P, that is $\sum_{P_i \in P, i=1,2\ldots,|P|} \sum_{e \in P_i} w(e)$.

To solve this problem, simply transforming into a max-flow problem is not applicable. However, it is observed that the edges for matching on the bipartite graph G''_w also correspond to the edges in the path cover on G''_w. As shown in Proposition 1, finding a maximum matching will lead to a minimum path cover, hence a minimum weight maximum matching on the graph will correspond to the minimum weight minimum path cover problem [28]. Consequently, we can transform the graph $G'_w(V'_w, E'_w)$ into a complete bipartite graph $G''_w(V''_w, E''_w)$ illustrated in Fig. 5(b), and instead solve a minimum weight maximum bipartite matching problem on G''_w. In G''_w, if the two trips are combinable, we use the same weight computed in Equation (4) on edges that connect v_i^d and v_j^o if the two trips are combinable. For all other edges, we assign the weight to be infinite (a sufficient large number M in actual implementation). The Hungarian method can again be used to solve the problem. By removing the matching edges that contain the weight of M from the result, we arrive at the final solution that corresponds to the minimum weight minimum path cover problem.

IV. Experiment Results

In this section, experiment results of the theoretical optimal system performance are presented for both optimal matching and trip integration. The results are carried out using the real world large-scale taxi data in NYC. The main idea of the experiment is to ensure all taxi trips are served at exactly the same time and location as in the data, meanwhile, find the optimal strategy that (1) minimizes the empty trip cost for all taxis (optimal matching), (2) finds the arrangement of minimum number of taxis with/without minimizing the empty trip cost (trip integration).

A. Results for Optimal Matching

Since the real world taxi data reveal an inefficient driver-passenger matching strategy, although optimal matching is applicable to real-world situations, it is not appropriate to directly implement the optimal matching on the observed data for a large time interval. This is because that the optimal matching rapidly fills up the passenger demand, leaving behind a system with excessive amount of unmatched taxis that significantly deviates from the observed taxi service system. In real world settings, these unmatched taxis could serve other potential trips, however, such trip information is not recorded in the data. Thus to evaluate the performance of the optimal matching, an experiment is carried out using actual taxi trip data from NYC on three short time periods (8:00–8:20, 12:00–12:20, 18:00–18:20) on Wednesday, Friday and Saturday of the tested week, which covers both the morning peak and evening peaks. The length of time interval ΔT is set to 5 minutes, which results in 4 consecutive time intervals in each time period.

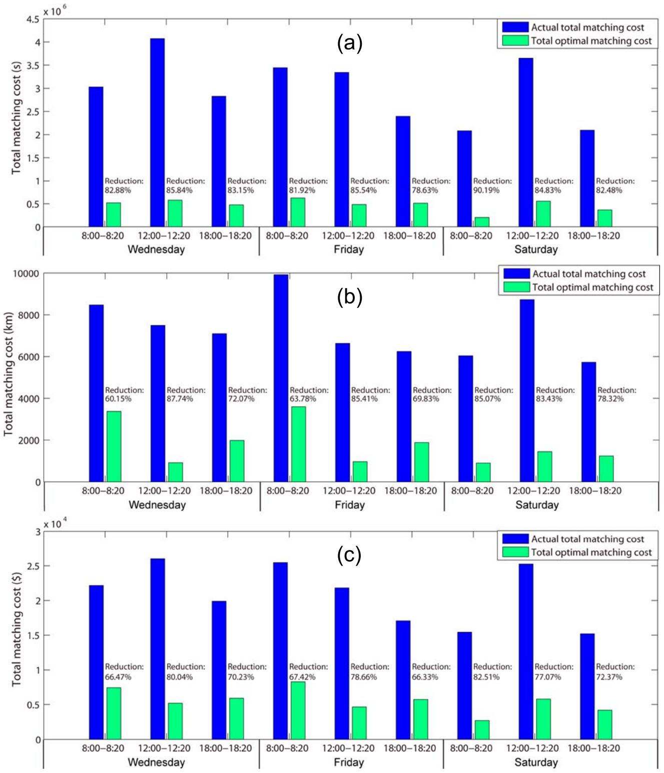
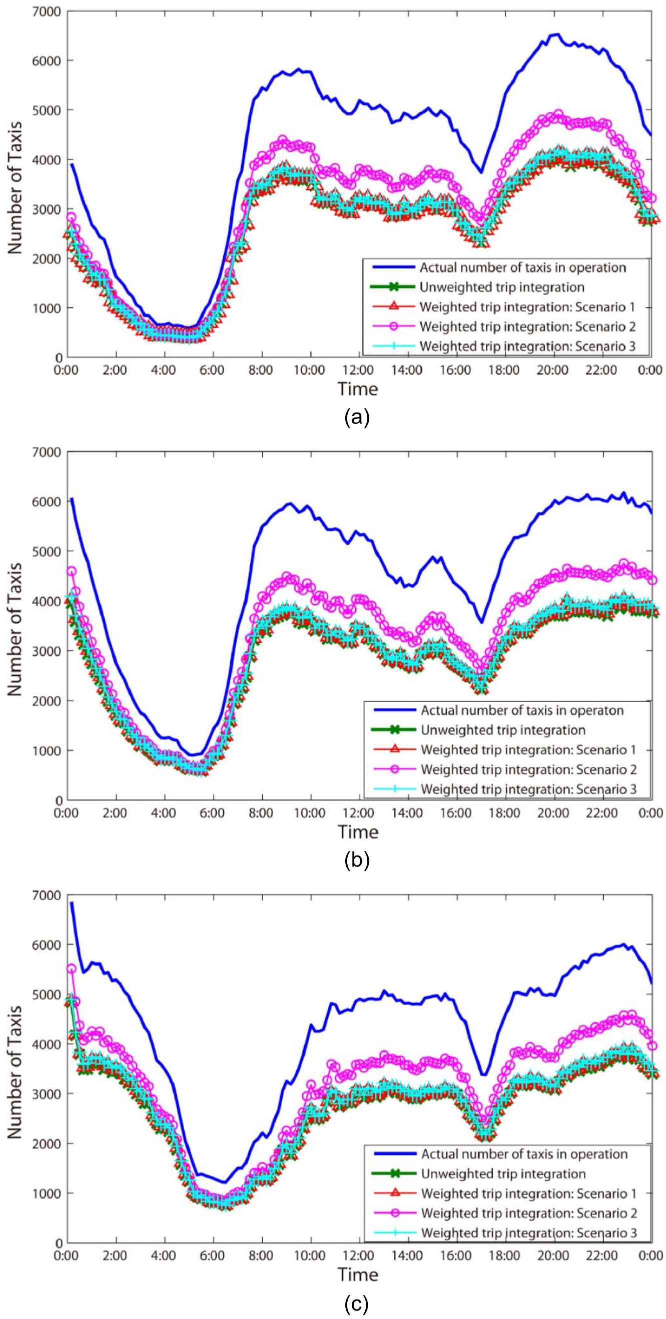


Fig. 6. Optimal matching results for different time of the day and days of the week (8:00–8:20,12:00–12:20, and 18:00–18:20 on Wednesday, Friday and Saturday). (a) Scenario 1. (b) Scenario 2. (c) Scenario 3.

The experiment results for optimal matching are presented in Fig. 6. From the optimal matching results, significant reductions are achieved in all time periods and scenarios. It is observed that the total taxi idle time (Scenario 1) can be reduced in the range of 78% to 90%, and the total empty trip distance (Scenario 2) can be reduced between 60% and 87%. For Scenario 3, in which the loss of revenue computation involves both the taxi idle time and empty trip distance, the reduction is observed to be around 66% to 82%. The results indicate that taxi drivers spend a significant amount of excessive time and travel distances in the road network looking for passengers, and a better scheduling and coordination for taxi-passenger matching can reduce taxi idle time and empty trip distance to a great extent. Apparently, the optimal matching strategy can greatly enhance the level of taxi service by reducing the waiting time for passengers and taxi drivers. The huge gap between the empty trips traveled in the actual taxi service system and the theoretical optimum uncovers the significant loss caused by the asymmetric information between taxi drivers and passengers.



(a)

(b)

(c)

Fig. 7. Number of taxis required to satisfy all trips observed in data. (a) Trip integration results for Wednesday. (b) Trip integration results for Friday. (c) Trip integration results for Saturday.

B. Results for Trip Integration

For the experiments on trip integration, we consider the length of time interval $\Delta T = 10$ min and test both unweighted and weighted trip integration on the entire selected Wednesday, Friday and Saturday data. Fig. 7 shows the number of taxis required to serve all trips observed in data for both unweighted and weighted cases in different scenarios. As expected, the unweighted trip integration yields the least amount of required taxis. The results show that for most cases, only 2/3 of the observed taxis are sufficient to satisfy all the trips in the data. The whole system output (served trips) can be greatly boosted if all taxis are fully utilized using an optimal trip integration strategy. For weighted trip integration, the results suggest that the number of taxis required for Scenario 1 and 3 are only slightly higher than unweighted case. The strategy that minimizes matching cost is advantageous to provide drivers with necessary information and achieves the greatest time saving during the period. While Scenario 2 requires more number of taxis, it is still capable of saving about 1/4 of total taxis. The higher number of taxis needed in Scenario 2 is probably due to the fact that taxi drivers are more likely to roam near the latest drop off location. When the idle time of the empty trip is not considered, the weighted trip integration is more likely to assign original taxi to each trip, which leads to lower performance compared with the other 2 scenarios.

The average matching cost from the actual taxi service system in NYC and the weighted trip integration are compared on the three tested days (Fig. 8). It is found that for all time periods and scenarios, the average matching cost from weighted trip integration is only about half of the actual average matching cost, even though fewer taxis are used to serve the same amount of trips in the optimal system. The reduction can be even more significant in some time periods such as from 10:00–18:00, when both taxi supply and passenger demand are high. The average matching costs of both the actual system and the system using weighted trip integration are much higher in late night (after 22:00) and early morning (before 6:00) compared with other time periods. This is mainly due to relatively fewer number of taxis and passenger trip demand during these time periods, which introduce unavoidable large empty trip cost (matching cost) for both actual system and the optimal system.

What is interesting lies in the difference of the results between the actual system and the optimal system beyond the previous time periods. Here we refer to the time periods as stable regions, which is 9:00–20:00 for weekdays and 10:00–21:00 for weekend. The stable regions are illustrated as the regions between the two vertical dash lines in Fig. 8. The average matching costs of weighted trip integration remain low and stable in all test cases, and converge to very similar values regardless of the tested days. For example, for Scenario 1, the average matching costs stabilize at around 85 s; for Scenario 2, they stabilize at 0.34 km and for Scenario 3, this value is about \$0.8. On the other hand, the actual matching cost fluctuates significantly, and has peaks around noon, when the number of taxi is also high in actual system. This observation reveals the existence of destructive competition in the taxi service system in NYC. Given the same set of taxi trips, the average trip matching costs in the stable region are observed to be stable in the perfectly coordinated taxi service system (weighted trip integration). The costs are barely affected by either the time of the day or the day of the week. However, in the real world system, the average matching costs oscillate drastically within the stable region. Without proper information, it is very likely that some drivers compete for limited passengers. The consequence is that only few drivers effectively win the business while the rest waste their time and fuels and have to start over the passenger search. The phenomenon is especially evident from 10:00–12:00 and 14:00–16:00, when there is less passenger demand in the network. The results indicate the

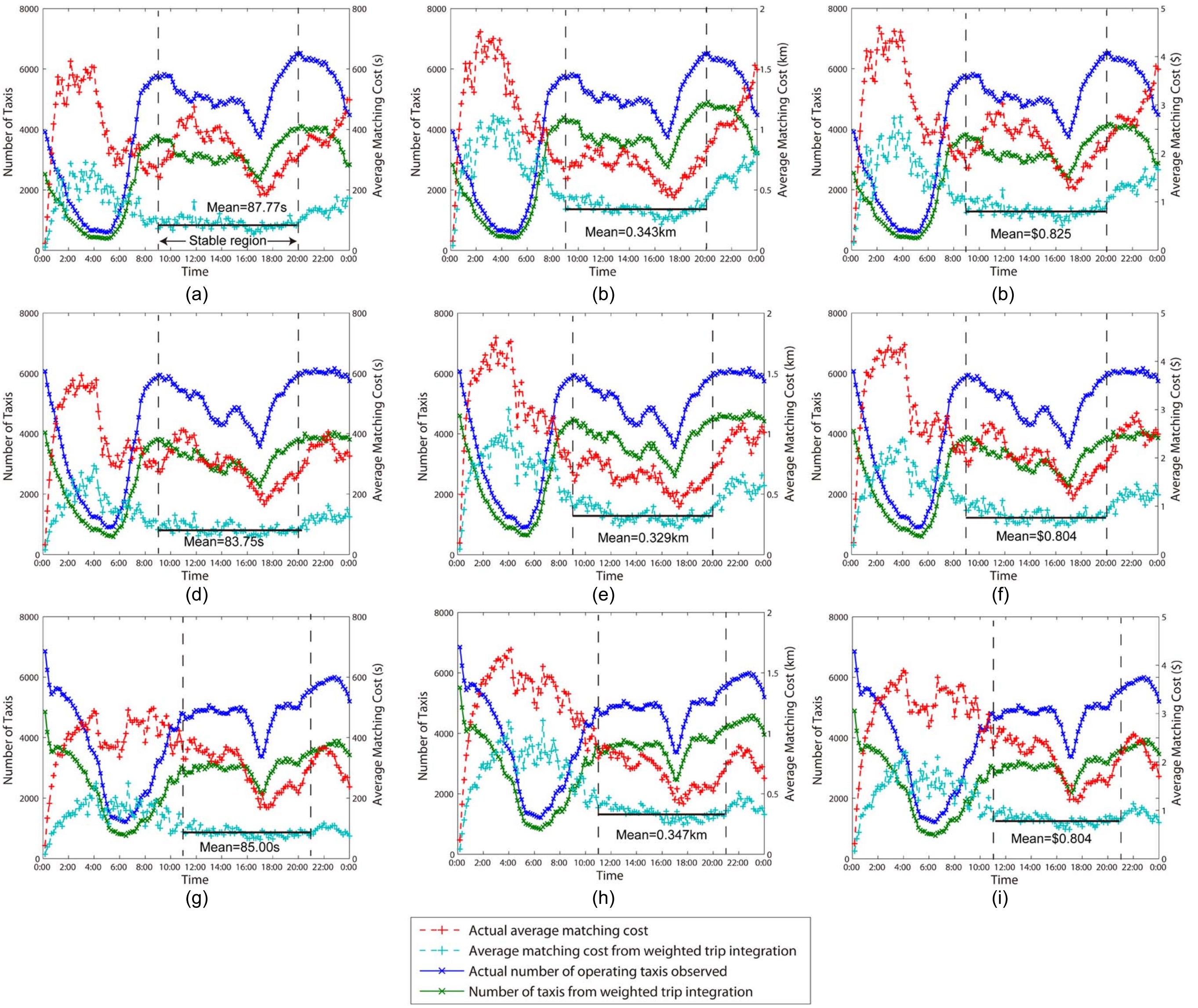


Fig. 8. Comparison of average matching costs for weighted trip integration. (a) Wednesday: Scenario 1. (b) Wednesday: Scenario 2. (c) Wednesday: Scenario 3. (d) Friday: Scenario 1. (e) Friday: Scenario 2. (f) Friday: Scenario 3. (g) Saturday: Scenario 1. (h) Saturday: Scenario 2. (i) Saturday: Scenario 3.

existence of destructive competition deteriorates the overall system performance of the current taxi service system in NYC.

C. Impacts of Choices on ΔT

As discussed previously, the length of the time interval ΔT in optimal matching and trip integration affects the performance of these two methods. Optimal matching generally requires a smaller ΔT while trip integration needs a larger one. From the optimization perspective, a larger ΔT will always lead to greater reduction of objective function value, since more trips are considered and optimized within each time period; from operation perspective, smaller value of ΔT is desired to avoid extra waiting time for users in each operation. To evaluate the impact and trade-off of different values of ΔT on optimal matching and trip integration, we conducted additional experiments using data from Wednesday. For optimal matching, ΔT values of 1 min, 2 min, 4 min, and 5 min are tested; whereas for trip integration, we tested larger ΔT values of 5 min, 7.5 min, 10 min, and 15 min.

The comparison of optimal matching results are presented in Fig. 9. In the figures, $\mathrm{TMC}_{\mathrm{opt}}$ refers the total matching cost from optimal matching, and $\mathrm{TMC}_{\mathrm{act}}$ refers the actual total matching cost. For each time period, $\mathrm{TMC}_{\mathrm{act}}$ is the same for different ΔT values, thus smaller $\mathrm{TMC}_{\mathrm{opt}}/\mathrm{TMC}_{\mathrm{act}}$ suggests greater reduction in total matching cost. As expected, the results show that larger ΔT indeed leads to greater reduction in terms of the total matching cost within each tested periods. It is also observed that when the time interval becomes very small $(\Delta T = 1)$, the reduction in total matching cost decreases greatly especially for Scenario 2, which suggests the time window might be too short to collect enough information for the full utilization of taxis. However, when moderate length of time interval is used $((\Delta T = 2,\ 4$ min), the differences in reductions of total matching cost are small among results for different ΔT values, which suggests the appropriateness of implementing optimal matching using even smaller time interval size.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

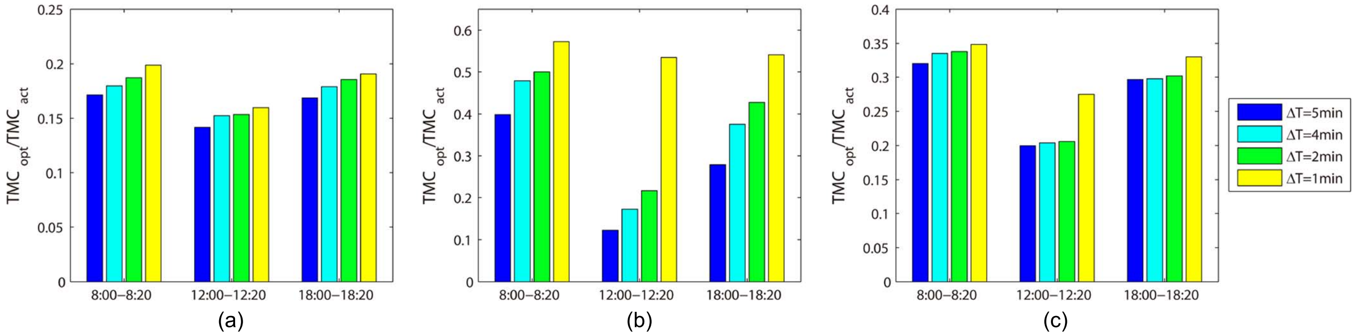


Fig. 9. Optimal matching costs for different values of ΔT on Wednesday. (a) Scenario 1. (b) Scenario 2. (c) Scenario 3.

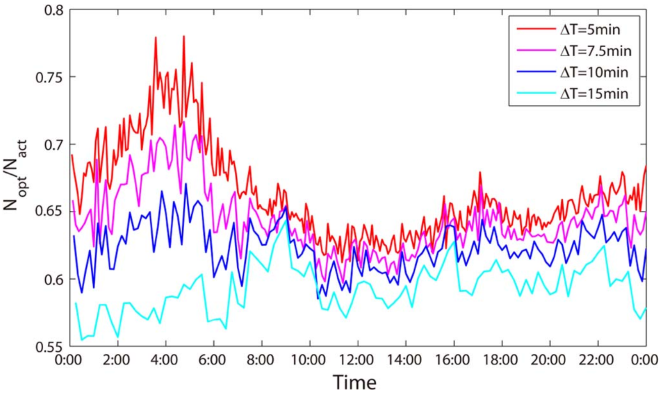


Fig. 10. Ratio of number of taxis required in unweighted trip integration versus actual number of taxi observed for different values of ΔT on Wednesday.

For results of trip integration, the ratio of number of taxis needed obtained from unweighted trip integration versus the actual number of taxi in operation is plotted in Fig. 10. In the figure, N_{opt} refers the total number of taxis obtained from unweighted trip integration in each time period, N_{act} refers the observed number of taxis in the actual system of the same time period. Smaller $N_{\text{opt}}/N_{\text{act}}$ indicates greater reduction in unweighted trip integration. The results confirm the intuition that greater ΔT value leads to more sufficient utilization of taxis (lower number of taxis needed). However, unlike optimal matching, the difference is larger in unweighted trip integration, since the maximum difference of number of taxi used for $\Delta T = 5\ \text{min}$ and $\Delta T = 15\ \text{min}$ can be as high as 20%. Although $\Delta T = 15\ \text{min}$ leads to the best unweighted trip integration performance, it is unrealistic to apply such long time interval in real world operation, whereas the $\Delta T = 10\ \text{min}$ tested in previous sections provide a reasonable consideration in balancing operation delay and optimization performance. The pattern of number of taxis needed from weighted trip integration is very similar to Figs. 7 and 10, and the average matching cost is similar to Fig. 8 regardless of the different value of ΔT. To avoid repetition, these results from weighted trip integration are not presented in this paper.

V. Conclusion

This paper presents the first study to quantify the efficiency level of the taxi service system in New York City using a real world large-scale taxi trip dataset. A hypothetical system-wide recommendation mechanism is assumed that allow both taxi drivers and passengers to share their trip information. Two approaches, namely optimal matching and trip integration are proposed to find the optimal strategy that minimizes the cost of empty trips, and the number of taxis required to serve all observed trips. The optimization problems in the two approaches are transformed into equivalent graph problems and solved using efficient polynomial time algorithms.

The results show that the optimal matching can reduce about 78% to 90% of the total taxi idle time, 60% to 87% of the total empty trip distances, and 66% to 82% of the total revenue loss of empty trips when different objectives are considered. For trip integration, the results show that in most cases, two third of all taxis are sufficient to satisfy all the trips observed in the data. For weighted trip integration, the actual average matching cost in terms of idle time, empty trip distance and revenue loss can be reduced to half even when fewer taxis are used. The existence of destructive competition in the taxi service system is also observed by comparing the actual system and the perfectly coordinated system governed by weighted trip integration. It also reveals the fact that in a perfectly coordinate system, the average matching costs remain almost the same in the stable region regardless the different time of the day or the day in the week.

The findings in this paper show that the actual taxi service system in New York City is far from efficient. The lack of sharing system-wide information between taxi drivers and passengers results in large amount of extra idle time and travel distances spent on empty trips. Moreover, if such a system-wide information sharing scheme exist, only 2/3 of the total taxis could be sufficient to serve all observed trips, the taxi shortage issue can be potentially resolved by better matching and integrating taxi trips. A further implication from the results is related to an important question that many cities face: should cities increase the existing number of taxis by adding new licenses, thereby potentially worsening congestion or encouraging adopting new technologies that enable centralized matching between taxis and passengers using globally shared information to improve system efficiency? This study provides an affirmative conclusion to the second option. Currently, this study only focused on analyzing the taxi service system for New York City, additional experiments can be conducted for other cities when similar datasets become available. This will further validate and show the applicability of the optimization approaches developed in this paper.

Our analysis also suggests an urgent need to adopt system level thinking into current taxi recommendation and dispatching system design. The current decentralized taxi recommendation systems that involve a subset of all drivers and passenger trips might benefit specific taxi drivers or passengers, however, may not necessarily improve the entire system performance. Some current taxi hailing apps can potentially make some taxi trips exclusive, which under certain situation will worsen the overall system performance. Future work can be done to extend the two approaches considered in this paper, and develop a taxi service recommendation and management system. To make the proposed algorithms work under real world situations, additional modules, such as path travel time estimation, and consideration of the possible imperfect information of future passenger demand need to be introduced. Furthermore, if the fairness among taxis needs to be ensured, the proposed approaches can still be used by adding a positive penalization term to the matching cost computation. The penalization term is a monotonically increasing function of the number of trips served by the taxi in the previous fixed length time period. Thus taxi drivers serving too many trips will be penalized and certain level of fairness can be guaranteed. All these aforementioned improvements will contribute to building a centralized taxi recommendation and management system that leads to a more efficient taxi service system and a more sustainable urban environment.

REFERENCES

[1] "New York City taxi and limousine commission 2012 annual report," New York City Taxi Limousine Comm., New York, NY, USA, 2012.

[2] "Transport—Hong Kong: The facts," Gov. HongKong, 2013.

[3] H. Yang, Y. W. Lau, S. C. Wong, and H. K. Lo, "A macroscopic taxi model for passenger demand, taxi utilization and level of services," *Transp.*, vol. 27, no. 3, pp. 317–340, Jun. 2000.

[4] T. Çetin and K. Yasin Eryigit, "Estimating the effects of entry regulation in the Istanbul taxicab market," *Transp. Res. A, Policy Pract.*, vol. 45, no. 6, pp. 476–484, Jul. 2011.

[5] C. F. Manski and J. D. Wright, "Nature of equilibrium in the market for taxi services," Transp. Res. Board Business Office, Washington, DC, USA, Tech. Rep., 1967.

[6] G. W. Douglas, "Price regulation and optimal service standards: The taxicab industry," *J. Transp. Econ. Policy*, vol. 6, pp. 116–127, 1972.

[7] C. F. Daganzo, "An approximate analytic model of many-to-many demand responsive transportation systems," *Transp. Res.*, vol. 12, no. 5, pp. 325–333, 1978.

[8] K. Wong, S. Wong, and H. Yang, "Modeling urban taxi services in congested road networks with elastic demand," *Transp. Res. B, Methodol.*, vol. 35, no. 9, pp. 819–842, Nov. 2001.

[9] X. Qian and S. V. Ukkusuri, "Spatial variation of the urban taxi ridership using GPS data," *Appl. Geograph.*, vol. 59, pp. 31–42, 2015.

[10] J. Aslam, S. Lim, and D. Rus, "Congestion-aware traffic routing system using sensor data," in *Proc. IEEE 15th Int. Conf. ITSC*, 2012, pp. 1006–1013.

[11] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi-passenger demand using streaming data," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1393–1402, Sep. 2013.

[12] X. Zhan, S. Hasan, S. V. Ukkusuri, and C. Kamga, "Urban link travel time estimation using large-scale taxi data with partial information," *Transp. Res. C, Emerging Technol.*, vol. 33, pp. 37–49, Aug. 2013.

[13] X. Zhan, S. V. Ukkusuri, and C. Yang, "A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data," *Autom. Construction*, doi:10.1016/j.autcon.2015.12.007, to be published.

[14] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, Sep. 2014.

[15] J. Lee, I. Shin, and G.-L. Park, "Analysis of the passenger pick-up pattern for taxi location recommendation," in *Proc. IEEE 4th Int. Conf. NCM*, 2008, vol. 1, pp. 199–204.

[16] Y. Ge *et al.*, "An energy-efficient mobile recommender system," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2010, pp. 899–908.

[17] K. Yamamoto, K. Uesugi, and T. Watanabe, "Adaptive routing of cruising taxis by mutual exchange of pathways," in *Knowledge-Based Intelligent Information and Engineering Systems*. New York, NY, USA: Springer-Verlag, 2008, pp. 559–566.

[18] N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-finder: A recommender system for finding passengers and vacant taxis," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 10, pp. 2390–2403, Oct. 2013.

[19] S. Ma, Y. Zheng, and O. Wolfson, "T-share: A large-scale dynamic taxi ridesharing service," in *Proc. IEEE 29th ICDE*, 2013, pp. 410–421.

[20] S. Ma and O. Wolfson, "Analysis and evaluation of the slugging form of ridesharing," in *Proc. 21st ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2013, pp. 64–73.

[21] P. M. d'Orey, R. Fernandes, and M. Ferreira, "Empirical evaluation of a dynamic and distributed taxi-sharing system," in *Proc. IEEE 15th ITSC*, 2012, pp. 140–146.

[22] X. Qian, X. Zhan, and S. V. Ukkusuri, "Characterizing urban dynamics using large scale taxicab data," in *Engineering and Applied Sciences Optimization*. New York, NY, USA: Springer-Verlag, 2015, pp. 17–32.

[23] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logics Quart.*, vol. 2, no. 1/2, pp. 83–97, 1955.

[24] N. Tomizawa, "On some techniques useful for solution of transportation network problems," *Networks*, vol. 1, no. 2, pp. 173–194, 1971.

[25] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency*, vol. 24. New York, NY, USA: Springer-Verlag, 2003.

[26] P. Santi *et al.*, "Quantifying the benefits of vehicle pooling with shareability networks," *Proc. Nat. Acad. Sci.*, vol. 111, no. 37, pp. 13 290–13 294, 2014.

[27] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, vol. 2. Cambridge, MA, USA: MIT Press, 2001.

[28] R. Rizzi, A. I. Tomescu, and V. Mäkinen, "On the complexity of minimum path cover with subpath constraints for multi-assembly," presented at the 4th Annual RECOMB Satellite Workshop Massively Parallel Sequencing, BMC Bioinformatics, Pittsburgh PA, USA, Apr. 2014.

Xianyuan Zhan received the B.E. degree in civil engineering from Tsinghua University, Beijing, China, and the M.S. degree in transportation engineering from Purdue University, West Lafayette, IN, USA, where he is currently working toward the Ph.D. degree in transportation engineering and the dual M.S. degree in computer science. His research interests include large-scale data analytics, complex networks, human mobility and activity pattern analysis, and transportation network modeling.

Xinwu Qian received the B.S. degree in transportation engineering from Tongji University, Shanghai, China, and the M.S. degree in transportation engineering from Purdue University, West Lafayette, IN, USA, where he is currently working toward the Ph.D. degree with the School of Civil Engineering. His research interests include big data analytics, complex network analysis, and transportation network modeling.

Satish V. Ukkusuri received the B.Tech. degree in civil engineering from IIT Madras, India; the M.S. degree in transportation systems from University of Illinois at Urbana Champain; and the Ph.D. degree in transportation systems from The University of Texas at Austin. He is currently a professor with Lyles School of Civil Engineering, Purdue University, West Lafayette, IN, USA. He is recognized nationally and internationally in the area of transportation network modeling and disaster management. His areas of interest include dynamic network modeling, large-scale data analytics, disaster management issues, and freight transportation and logistics. He has published extensively on these topics in peer-reviewed journals and conferences.