Minimum fleet algorithm considering human spatiotemporal behaviours

Zhi-Dan Zhao[a,b,1,*], Yu Wang[a,b,1], Wei-Peng Nie[c,d,e], Shi-Min Cai[c,d,e], Celso Grebogi[f]

a. Complexity Computation Lab, Department of Computer Science, School of Engineering, Shantou University, Shantou 515063, China.

b. Key Laboratory of Intelligent Manufacturing Technology (Ministry of Education), Shantou University, Shantou 515063, China.

c. Compleχ Lab, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China.

d. Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China.

e. Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 610054, China.

f. Institute for Complex Systems and Mathematical Biology, School of Natural and Computing Sciences, King's College, University of Aberdeen, Aberdeen AB24 3UE, UK.

Abstract

With the development of information technology, more and more travel data have provided great convenience for scholars to study the travel behavior of users. Planning user travel has increasingly attracted researchers' attention due to its great theoretical significance and practical value. In this study, we not only consider the minimum fleet size required to meet the urban travel needs, but also consider the travel time and distance of the fleet. Based on the above reasons, we propose a travel scheduling algorithm that comprehensively considers time and space costs, namely, the Spatio-Temporal Hopcroft-Karp (STHK) algorithm. The analysis results show that the STHK algorithm can not only significantly reduce the off-load time and off-load distance of the fleet travel by as much as 81% and 58% respec-

[*] Correspondence should be addressed to Zhi-Dan Zhao (zzhidanzhao@gmail.com) and Celso Grebogi (grebogi@abdn.ac.uk)

[1] Contributions: Z.-D. Zhao and Y. Wang contributed equally to this work.

tively, but also retains the heterogeneous characteristics of human travel behavior. Our study indicates that the new planning algorithm can not only provide the size of the fleet to meet the needs of urban travel, but also reduce the waste of travel time and distance, thereby reducing energy consumption and reducing carbon dioxide emissions. Concurrently, the travel planning results also conform to the basic characteristics of human travel, and have important theoretical significance and practical application value.

Keywords: Human travel behaviour, Minimum fleet, Spatio-Temporal Hopcroft-Karp

1. Introduction

The study of human travel behavior has received greater attention, especially its important impact in solving traffic congestion, disease prevention and control, and environmental protection and emission reduction [1, 2, 3, 4]. With the acceleration of urbanization, urban travel is becoming more and more convenient, which also brings many problems such as traffic congestion, carbon dioxide emissions, and air pollution [5, 6, 7, 8, 9]. A major current focus in urban travel is carpooling, which is older than horse-drawn carriage travel, and recent innovations have made carpooling easier, more convenient, and more efficient than ever. Innovative travel services based on sharing can reduce travel costs, alleviate congestion, and reduce greenhouse gas emissions, and have huge economic, social, and environmental benefits [10]. Carpooling services represented by taxis are changing the mode of transportation in cities, due to their multi-body and self-organizing characteristics [11], providing timely and convenient transportation services for anyone, any place, and any time. These services also have huge potential positive social impacts in terms of pollution, energy consumption, and traffic congestion [12, 13].

On the one hand, a lot of research has focused on the demand for carpooling services and dynamic pricing, and has achieved very meaningful results. Çolak et al. [14] combined the road network of five different cities with the travel demand curve during morning peak hours and found that the dimensionless ratio of road supply to travel demand explained the percentage of congestion time lost. Boesch et al. [15] found that the relationship between service demand and required fleet size is non-linear, and the ratio increases as demand increases. If the waiting time is extended, it is possible to reduce the number of vehicles. Storch et al. [16] modelled the basic incentives behind individual carpooling decisions and found that there are two opposing mechanisms for carpooling decisions: constant or decrease as demand increases. Combining game theory and time series analysis, Schröder et al. [17] revealed how and under what conditions dynamic pricing encourages online ride-hailing drivers to act collectively, leading to abnormal supply shortages. Molkenthin et al. [11] observed an universal scaling law and found that one scaling parameter could jointly captures the influence of network topology and demand distribution.

On the other hand, the route planning of carpooling services has also received a lot of research attention. Traditionally precise methods based on mathematical planning (traditionally used in traffic system design) can only handle a few thousand trips or vehicles at most [18, 19, 20]. Alonso-Mora et al. [21] provided a general mathematical model for real-time high-capacity ride-sharing, which could be extended to a large number of passengers and trips, and could dynamically generate the best routes considering online demand and vehicle locations. Especially recently, with the rapid development of transportation technology and communication technology, the management and planning of taxis have become more systematic and technological. For instance, in order to answer the question of the minimum number of taxis needed to meet the travel needs of the city, a few works addressed the "minimum fleet problem". Vazifeh et al. [22] processed the actual itinerary data into a multi-segment itinerary and stored it in the form of a bipartite graph, and determined the minimum fleet size through the Hopcroft-Karp algorithm [23] for solving the maximum matching of the bipartite graph, and gave a travel plan for the taxi fleet. Additionally, Fagnant et al. [24] studied the dynamic ride sharing (DRS) on the shared autonomous vehicle (SAV) network in Austin, Texas, and observed that DRS would reduce the average service time and travel cost of SAV users. Ke et al. [25] proposed a deep learning method that considered spatial dependence, temporal dependence and exogenous dependence at the same time, and it captured the correlation between spatio-temporal features.

Although the above methods can give the minimum number of fleets that meet the travel demand, or consider the time and space consumption of the fleet, the minimum fleet algorithm that integrates the minimum spatio-temporal consumption is rarely mentioned. This work comprehensively considers the travel time and space consumption, and proposes an algorithm that integrates the vehicle's empty distance and the driver's tolerance time. The experimental results show that our proposed algorithm can greatly reduce the time consumption and distance waste of taxis, and the human travel behavior characteristics of the algorithm results are also consistent with our everyday life. It shows that on the basis of saving time and distance consumption, our method is also in line with the user's daily basic travel habits. Then, this paper is organized as following. The section 2 is the data processing, the algorithm principles, and implementation steps, such as specific optimization processing details; the section 3 is the analysis of experimental results. This work compares and analyzes the original data and the improvements of the results of the algorithm; the last section 4 presents the summary and conclusions, including possible future development directions and a discussion of related applications and extensions.

2. Materials and Methods

This section introduces mainly the details of the data set and its processing, as well as the description of related algorithms. Firstly, we introduce the raw data and how to incorporate the raw Global Positioning System (GPS) location information into itineraries. Secondly, the original maximum matching algorithm for bipartite graphs, the Hopcroft-Karp algorithm, is introduced. Then, we describe the Spatio-Temporal Hopcroft-Karp algorithm, an algorithm proposed in this work, to analyze how to reduce the off-load distance and off-load time. Finally, the Gini coefficient that quantifies the heterogeneous characteristics of human travel behavior is presented.

Data collection

The data set used in this work is the travel record data of Chengdu taxis for about one month. We divide the data into daily data sets, each of which is about 55 million records (as shown in Table 1). Each recorded data item includes anonymous vehicle ID, GPS coordinates, and vehicle status sign (on-load or off-load, as demonstrated in [26, 27]). We define the on-load mode as the process in which the vehicle status sign changes from 0 to 1, and then to 0. The off-load mode is that the vehicle status sign changes from 1 to 0, and then to 1.

The detailed process of converting GPS location information to itinerary information is as follows. Firstly, the daily GPS records is sorted in ascending order of time. Secondly, the starting position of the vehicle, anonymous vehicle ID, timestamp, and the vehicle status sign is recorded. Thirdly, each piece of GPS information is scanned to determine whether the vehicle status sign has changed. If it changed, which indicates the end of one itinerary, the information of the itinerary is recorded. Finally, the obtained itinerary records are distinguished according to the vehicle status sign, e.g., for the off-load data (the vehicle status sign is 0), the actual distance of the itinerary needs to be recorded, and for the on-load data (the vehicle status sign is 1), the service time and travel distance of the itinerary is calculated. In order to filter out noise data, the itinerary with a service time of less than 1 minute or more than 240 minutes, and distance less than one kilometer is deleted, and the operation process is similar to the previous studies [26, 27]. Inevitably, this operation reduces some GPS information, but it is necessary to ensure the validity of the itinerary. After the above process, approximately 240,000 itineraries that meet our conditions are extracted every day, and the number of effective taxis exceeds 13,000 every day. Detailed data information is shown in Table 1.

Hopcroft-Karp algorithm

Hopcroft-Karp algorithm is an algorithm that takes bipartite graphs as input and produces maximum cardinality matching as output, which was independently discovered by John Hopcroft and Richard Karp [23] and Alexander Karzanov [28]. Bipartite graph (G) is one in which the vertices can be divided into two set U and V and the edges of the graph pass between the two sets but not within the same set [29]. In order to make the subsequent paragraph easier to read, we have defined the symbols used in this work and their corresponding explanations in Table 2.

The Hopcroft-Karp algorithm repeatedly increases the size of partial matches by finding an augmenting path, which is a set of edges that starts at an unmatched vertex and ends at another unmatched vertex. Additionally, the edges in the path alternate between being in the matching and out of the partial matching. The algorithm runs in phases and consists of the following steps. In the beginning, breadth-first search is used to find an augmentation path. It partitions the vertices of the graph into layers of matching and not matching edges. For the search, it starts from a free node in U, which forms the first layer of the partitioning. The search ends at the first layer k, where one or more free nodes in V are reached. Secondly, the free nodes in V are added to a set called F. This means that any node added to F is the ending node of an augmenting path—and a shortest augmenting path at that since the breadth-first search finds shortest paths. Thirdly, once an augmenting path is found, a depth-first search is used to add augmenting paths to current matching M. At any given layer, the depth-first search follows edges that lead to an unused node from the previous layer. Paths in the depth-first search tree must be alternating paths. Once the algorithm finds an augmenting path that uses a node from F, the depth-first search moves on to the next starting vertex. Finally, the algorithm terminates when the algorithm can find no more augmenting paths in the breadth-first search step.

Spatio-Temporal Hopcroft-Karp algorithm

In order to minimize energy consumption, reduce carbon dioxide and other emissions, and reduce environmental pollution, a feasible goal is to consider the time and space consumption of users' daily travel [12, 10, 13, 9, 6, 5]. In order to achieve this goal, the aim is to reduce unnecessary travel distance and waiting time, for example, to reduce the off-load time and off-load distance of taxis. Based on the original algorithm, this work puts the overall operating cost of the fleet into consideration. For example, the total off-load trip is added as the travel cost for comprehensive evaluation. At the same time, we consider the waiting time and empty travel distance of the driver to optimize the itinerary of the fleet, while ensuring a reduction in the size of the fleet, and minimizing the time and space consumption of the fleet. Based on this, we propose the Spatio-Temporal Hopcroft-Karp (STHK) algorithm, as shown in Figure 1, which is implemented as follows.

First, the itinerary that meets the conditions is selected and formed into a travel trajectory network. In the process of constructing the network, on-load itineraries are added to the travel trajectory network, and off-load itineraries that meet certain conditions are also added to the network. This condition is the connection time between the two trips, that is, the customer's tolerance time δ. This concept is similar to that in [22]. Here, δ is selected as 8 minutes. Figure 1 (a) is a schematic diagram of the travel trajectory network. The circles of different colors are the starting or end points of the different itineraries, the solid black arrow is the on-load itinerary and direction (from start point to end point), and the purple dashed arrows are the off-load itinerary and direction. The circle number on the black arrows are the on-load itinerary number, and the red number on the purple arrow is the weight of the off-load itinerary.

The second step is to construct a bipartite graph network for subsequent calculations based on the itinerary network in Figure 1 (a). The specific construction process is exactly the same as in [22]. That is to say, all the on-load trips are sequentially converted into the nodes of the two branches of the bipartite graph and then connected, and all possible off-load trips are also connected in turn. For example, on-load itinerary 1 may be connected to other on-load itineraries 2, 3, and 6, through the off-load itineraries (indicated by the red dotted lines) that meet the conditions. In this work, we use the Euclidean distance of the GPS position of the start and end points as the weight of the off-load itinerary. Although we only consider distance as the weight, taking into account the proportional relationship between distance and time, as shown in Supplementary Information Figure S1. We find this processing method to be reasonable. In addition, future research should more accurately consider the GPS real road distance of the start and end locations as the weight.

The third step, as shown in Figure 1 (c), is the process of finding the maximum cardinality matching in the bipartite graph, which is similar to the HK algorithm [23]. However, the original HK algorithm does not consider the weight of the bipartite graph, while the STHK algorithm considers the spatio-temporal characteristics of the trip as the weight. For example, in each matching, if there are multiple candidate connected edges (that is, conflicts), the connected edge with the highest weight is given priority to be the connected edge, as shown by the red line in Figure 1 (c). Moreover, compared with the maximum matching algorithm Kuhn-Munkres (KM) [30, 31], which considers the weight, the major difference is that the STHK algorithm only considers the weight when the candidate connected edges are in conflict, while the KM algorithm considers the total weight of the node to be optimal.

Finally, according to the above-mentioned STHK algorithm process, we get a new itinerary plan for urban taxis. The planning result is shown in Figure 1 (d), where the number of arrows in different colors indicates the size of the required fleet, and the edges of the arrows of the same color indicate the re-planned itinerary required by the same vehicle. Of course, due to the limitation of data, we only got the global static plan; future research should focus on the local dynamic plan.

Gini coefficient

The Gini coefficient [32] was proposed by Gini to measure the degree of inequality in a distribution. For example, in economics, it is most commonly used to measure the degree of deviation between a country's wealth or income distribution and an equal standard distribution. A Gini coefficient of 0 means exactly the same and no different distributions (for example, everyone has the

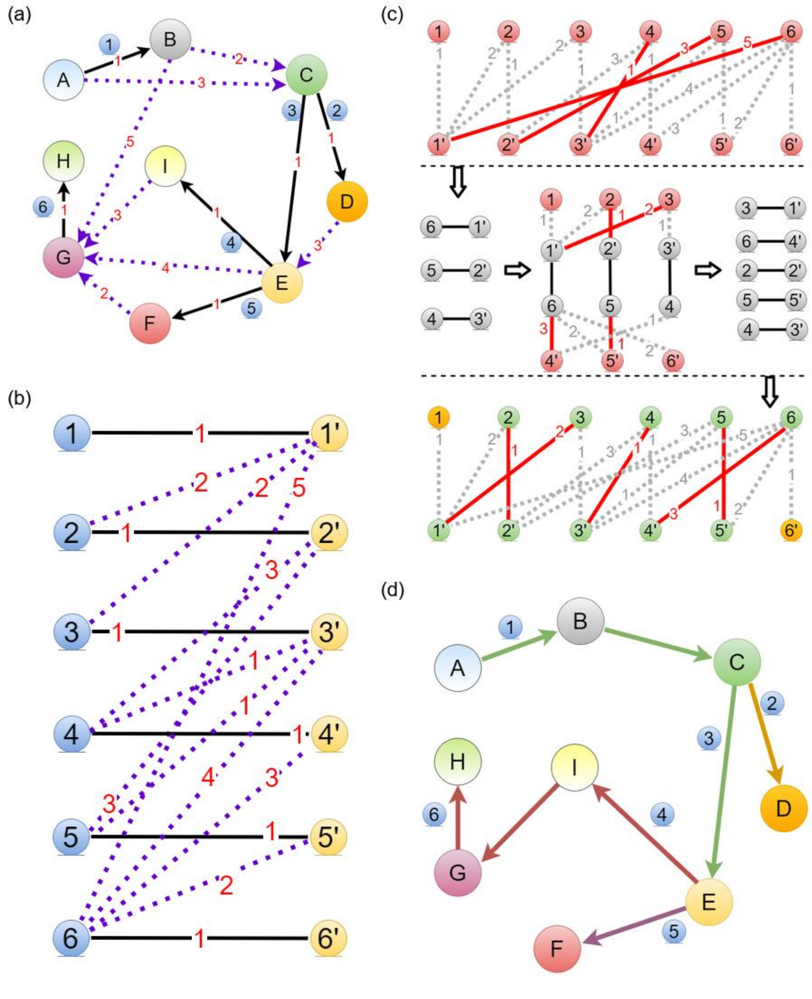

Figure 1: **The illustration of Spatio-temporal Hopcroft-Karp algorithm**. **(a)** Multiple travel requests are abstracted into a directed network, where the black solid arrow represents the user's real needs (or on-load process [26, 27], the need must be met) from the starting posiion to the destination, and the red dashed arrow represents the taxi off-load itinerary [26, 27]. The red number on the arrow indicates the cost of the itinerary. **(b)** This figure indicates that the itinerary graph 1 (a) is converted into a bipartite graph. The black solid lines indicate the connection of on-load itinerary, the purple dashed lines indicate that the possible itinerary is connected together, and the red number on the line segment indicates the cost of the itinerary. **(c)** The calculation process of the STHK algorithm is shown here, where the solid red line represents the path selected by STHK. **(d)** The planned taxi route plan is recalculated by the STHK algorithm. The connecting edges of different colors indicate the number of taxis after re-planning, and the connecting edges of the same color indicate the itinerary completed by the same taxi.

same income). A Gini coefficient of 1 indicates the largest inequality between the distribution values (for example, for an individual who has all income or consumption, but the other individual has not any income or consumption). One method is to define the Gini coefficient as half of the relative average absolute difference, which is mathematically similar to the definition of the Lorentz curve [33]. The average absolute difference is the absolute difference of all item pairs, and the relative average absolute difference is the normalized ratio by dividing the average absolute difference by the average $\bar{x}$. If x_i is the occurrence frequency of taxis in area i and there are N areas in total, the calculation formula of the Gini coefficient is given by the following formula 1:

$$G = \frac{\sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} |x_i - x_j|}{2N^2 \bar{x}} \tag{1}$$

3. Results

Here, we analyze the itineraries of Chengdu taxis, a total of about 6.8 million itineraries, and we plan the route of these itineraries in days according to the STHK algorithm. We compare the original data and the performance results of the STHK algorithm optimized travel in terms of distance and time. Remarkably, our results show that the STHK algorithm can reduce the off-load distance by 58% and the off-load time by 81%. Meanwhile, we analyze the heterogeneity of the distribution of the number of occurrences and the distribution of jumps in different regions, as well as the Gini coefficient of these two quantities over time. It is found that the average Gini coefficient is extremely heterogeneous regardless of the region or even the transfer between different regions. In short, the results of our quantitative analyses demonstrate that the STHK algorithm performs well in terms of space and time, and of heterogeneous human behavior.

Overall optimization of time and space.

Previous studies have mostly focused on the size of the urban taxi fleet, and have achieved good results, which could reduce the size of the taxi fleet [22, 34, 35]. According to our proposed STHK algorithm, impressive optimization results of the travel system in terms of distance and time resources are obtained.

Figure 2 (a) displays the comparison results of the actual and the total empty distance optimized by the STHK algorithm. From this result, we observe that the total off-load distance has dropped sharply from 21,303,835 kilometers to about 9,043,792 kilometers, a drop of 58%. However, this result simply considers the Euclidean distance between the start point and the endpoint, and future research should focus on the actual road travelled distance. Our result clearly demonstrates that the route arrangement re-planned by the STHK algorithm can not only ensure the normal travel demand of the city, but also significantly reducing the empty distance of the actual travel, thereby considerably reducing the consumption of fossil and the waste of energy and consequently, considerably reducing carbon emissions and protecting the environment. Similarly, we compare the actual data with the off-load time of the trip after the STHK algorithm re-planned, and the result is shown in Figure 2 (b). It indicates that the total off-load time after the re-planning of the STHK algorithm has been significantly reduced, from 4,609,332 hours to 863,222 hours, a drop of up to 81%. This result indicates that the STHK algorithm has a striking effect in saving the total cost of travel time in the city and can alleviate the congestion in the city.

The fluctuation of the itinerary in the city is a particularly noticeable problem [36]. The same road or intersection has obvious differences at different times of the day (working hours and off-working hours). There are also great differences between different days (holidays and working days) [26, 27]. Figure 2 (c) shows the daily fluctuations of the off-load travel distance. It can be seen from the figure that the daily off-load distance of the original trip data fluctuates greatly, from the lowest 678,787 kilometers to the highest 845,611 kilometers, with an average of 760,851 kilometers and a standard deviation of 36,864. Although this result is the distance change of the off-load trip, it can also clearly indicates traffic travel fluctuations. Surprisingly, the STHK algorithm's re-planned trip has an average daily off-load journey of 322,993 kilometers, which is about 58% lower than the actual data. In addition, the range of fluctuation is also significantly reduced, from 312,716 to 329,805 kilometers, the standard deviation is only 4,114. It is further confirmed that the STHK algorithm can dramatically reduce the off-load travel distance. This result implies that the STHK algorithm can not only greatly reduce energy consumption, reduce exhaust emissions, and protect the environment, but also reduce fluctuations in urban travel demand, thereby helping to keep daily travel relatively stable.

Daily off-load time is a very important indicator that reflects the daily

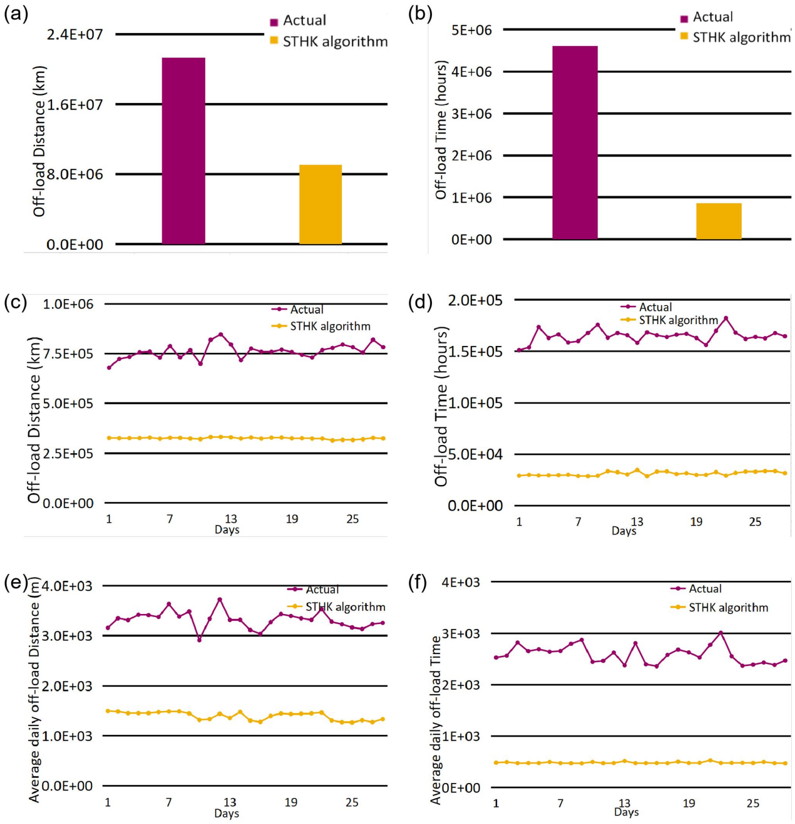

Figure 2: **The results of STHK versus Actual Data**. **(a)** The total off-load distance of actual data versus the STHK algorithm. The ordinate axis represents the total off-load distance in kilometers. **(b)** The total off-load time of actual data versus the STHK algorithm. The ordinate axis shows the total off-load time in hours. **(c)** The trend of the daily off-load distance, in which the X-axis represents the time in days, and the Y-axis represents the empty distance in kilometers. **(d)** The change trend of daily off-load time, in which the X-axis is time, and the Y-axis demonstrates the off-load time in hours. **(e)** The trend of the average distance of one off-load trip of a taxi. The horizontal axis represents the day of the month, and the vertical axis displays the average distance of one off-load trip, the unit is meter. **(f)** The trend of the average off-load time. The horizontal axis represents the day of the month, and the vertical axis displays the average off-load time, the unit is second. The actual results and the STHK algorithm results are represented by purple and yellow histograms and line segments, respectively.

work efficiency and work intensity of taxi drivers. Similar to the analysis of off-load distance, we compare the actual data and the daily off-load time processed by the STHK algorithm. Figure 2 (d) displays the fluctuations of the daily off-load time. The actual data has a large fluctuation range of the daily off-load time, from the minimum of 150,802 hours to the maximum of 182,016 hours, with an average value of 164,619 hours and a standard deviation of 6,407. The off-load time of the itinerary re-planned by the STHK algorithm fluctuates from 28,392 hours to 34,437 hours, with an average value of 30,829 and a standard deviation of 1,879. The average off-load time has been reduced by 81%. Such a large reduction shows that the STHK algorithm can greatly reduce the off-load time of taxis, and the daily off-load time fluctuations are also smaller. This may be due to the fact that the vehicles in the STHK algorithm continuously provide services to passengers, and the next trip destination is determined. Therefore, compared with the actual data, it is more efficient to find passengers without a goal, and the empty time consumed is greatly reduced. Future research should pay more attention to taxi scheduling and forecasting in real situation. In general, the results of the STHK algorithm are very encouraging, indicating that there is still a lot of room for improvement in the service of urban taxi travel.

For one taxi, the off-load distance and off-load time required to carry a passenger are two important economic and efficiency indicators. How to reduce the off-load distance and off-load time is a problem of important theoretical significance and application value. Here, we compare the average off-load distance and off-load time spent by each taxi to obtain one passenger under actual and STHK algorithm optimization, respectively. Figure 2 (e) shows the average off-load distance of each taxi per day and the average off-load distance optimized by the STHK algorithm. The average off-load distance represents the average value of off-load trip. The actual data's off-load distance ranges from 2,903 meters to 3,721 meters, while the STHK algorithm's off-load distance ranges from 1,261 meters to 1,473 meters, a reduction of up to 58%. Meanwhile, we can observe that the average off-load distance obtained by the STHK algorithm has smaller fluctuations than the actual off-load distance. Figure 2 (f) compares the average time of vehicles per day in Chengdu and the average time of the STHK algorithm. In the actual data, the average time of a single off-load varies from 2,353 seconds to 3,008 seconds, which means that it takes more than 40 minutes for a taxi to find the next passenger, while the average off-load time of the STHK algorithm is reduced to about 480 seconds. That is, it only takes eight minutes to serve the next passenger, reducing the off-load time by 81%. This result suggests that the endurance time δ, the waiting time to load a passenger, is a very effective influencing factor, and future research should pay more attention to the impact of endurance time δ. It can be seen that the actual taxi spends much time to search for the next passenger, and the time spent waiting at a specific location occupies a considerable proportion. Future research should focus on reducing these unnecessary waiting times and improving the efficiency of taxi operation, thereby reducing the total cost of taxi travel distance and time.

Analysis of influencing factors

In order to measure the influence of the two main factors, the tolerance time (No improvement 1) and vehicle conflict (No improvement 2) in the STHK algorithm, we conduct relevant experiments in this section to compare and analyze the differences in the results of the two factors. The comparison process is to remove one of the two factors each time, and then statistically analyze the experimental results after removing the factor, so as to evaluate the influence of the factor. Figure 3 displays the optimization results of the above two factors in the off-load distance and time of the STHK algorithm, respectively.

Firstly, we analyze the optimization results of the STHK algorithm's two factors for the off-load distance. If the impact of tolerance time is not considered, the result is shown as the yellow line in Figure 3 (a). We observe that the average off-load distance has dropped from 3,312 meters to 1,554 meters, which is 53% lower than actual data and verifies the optimization of the endurance time. In addition, the cyan line demonstrates the off-load distance is 2,569 meters when vehicle conflicts is not considered, which is 22% lower than actual data. This result imply that the vehicle conflicts has a more significant impact on the off-load distance.

Secondly, the impact of the two factors of the STHK algorithm on the off-load time consumption was shown in Figure 3 (b). We find that when the driver's tolerance time is not considered, the average off-load time is about 882 seconds, which is 1701 seconds less than the actual data (a drop by 66%), indicating that it shows optimization effect on the off-load time. Moreover, when the vehicle conflict is not considered, the off-load time is reduced from 2,583 seconds to 534 seconds, indicating that this factor has the greatest impact on the off-load time. Our experimental results demonstrate that considering the endurance time have basically effect on the average daily

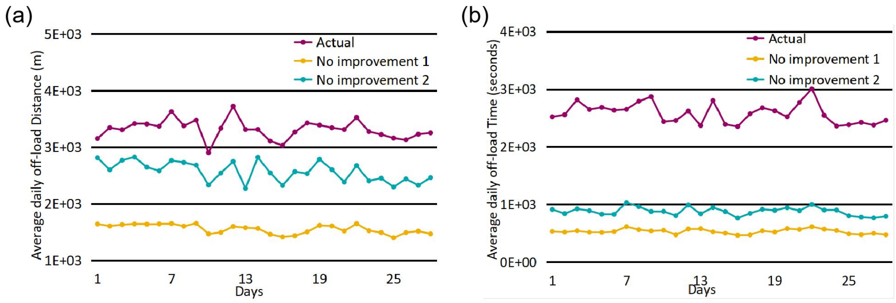

Figure 3: **Analysis of the effectiveness of the two main improvement factors**. **(a)** the average off-load distance of taxis in Chengdu in a month and the change of the average off-load distance of the STHK algorithm after removing one of the improvement factors. The abscissa represents days in the month, and the ordinate represents the average off-load distance of taxis (the unit is meters). **(b)** The trend chart of the average off-load time of taxis in one month in Chengdu and the average off-load time of the STHK algorithm after removing one of the improvement factors. The abscissa represents days of the month, and the ordinate represents the average off-load time of taxis (the unit is seconds). The actual off-load result is represented by the purple line, the result of no tolerance time (No improvement 1) is represented by the yellow line, and the result of no vehicle conflict (No improvement 2) is represented by the cyan line.

off-load time. However, ignoring the influence of vehicle conflict inventory factors is the most effective of the above mentioned factors. Furthermore, the experimental results confirmed that the vehicle conflict judgment factor is the most influential factor on the STHK algorithm, whether it is viewed from the time or space level. Our experimental results convincingly confirm that these two factors we proposed that drivers endure time and vehicle conflicts can optimize taxi travel from space to time.

Our experiments mainly test the optimization effect under the action of different factors. The experimental results show that when vehicles have conflict, determining the matching edge by calculating the conflicting vehicles' off-load time can yield better optimization results. In conflict judgment, increasing the driver's endurance time makes the vehicle path choice-less but increases the off-load time. These results indicate that the taxi service platform system's future research and design should focus on the influence of the driver's endurance time, vehicle conflict, and the most suitable strategy should be selected according to the actual situation.

Heterogeneity distribution of taxi trips

The heterogeneity of the distribution of taxi trips was the main focus in previous studies [37, 38, 39, 1, 26, 27]. Here we compare and analyze the travel distribution characteristics of actual data and STHK algorithm planning. Figure 4 shows the distribution characteristics of travel demand, the jump characteristics between different regions, the distribution of different regions, and the Gini coefficient of jumps between regions.

It is known from the literature that users' travel pattern obeys a heterogeneous distribution [37, 38, 26, 27]. Figure 4 (a) shows the travel characteristics of the actual data. In general, from which it can be inferred that there is a strong heterogeneity in travel demand in different regions, with average daily travel demand values varies from 469 to 60,027. It can be observed from Figure 4 (a) that not only the travel demand between regions is dramatically different, but the travel demand within the same region is also obviously uneven. The results are qualitatively similar to those of earlier studies [26, 27]. It should, however, be noted that the heterogeneity of these heterogeneous distributions may be caused by some natural environments in different regions, such as lakes and parks, etc. This phenomenon shows that the user's taxi behavior has a preference feature similar to other human mobile behaviors [37, 38, 39, 1, 3, 26, 27]. Similarly, Figure 4 (b) shows the taxi distribution characteristics after the algorithm STHK is re-planned. Surprisingly, there is a substantial similarity between the STHK algorithm results and the actual data, indicating that our optimization algorithm is in line with human taxis' actual travel behavior. Comparing Figure 4 (a) and Figure 4 (b), it can be seen that although the STHK algorithm has drastically changed the details of the matching process, it still retains the actual human preference characteristics, which implies that the optimization results of our algorithm does not make users feel inconvenient.

Human transfer behavior among different nodes (areas) is an important indicator, and it has a non-negligible influence in human behavior modelling and infectious disease model construction [37, 38, 39, 2, 3, 27]. Figure 4 (c) shows the jump characteristics between different areas in the actual data. We observe that there is a strong heterogeneity in the transfer of taxis in different blocks. The average of the largest transfer ratio is 0.210 and the average of the smallest transfer ratio is 0.004. The difference between the two is up to 55 times. The transfer feature after the re-planning of the STHK algorithm is shown in Figure 4 (d). It can be seen that regardless of the proportion of jumps in the same region or the proportion of jumps in different

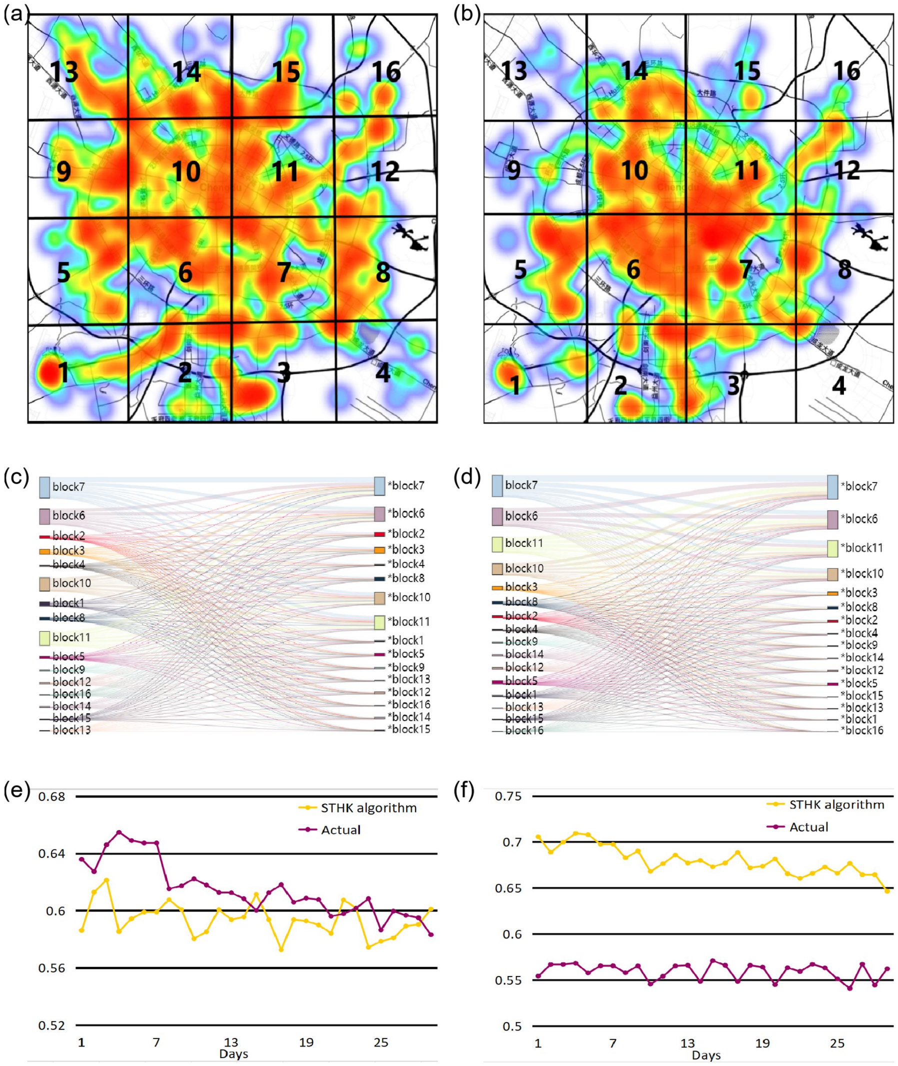

Figure 4: **The characteristic distribution of taxi travel**. **(a)** the actual distribution heat map of taxi travel records. The entire Chengdu area is divided into 16 blocks. **(b)** The heat map of the taxi distribution after the re-planning of the STHK algorithm. **(c)** The Sankey diagram of the actual flow of taxis. The flow of taxis from the departure area on the left to the destination area on the right. **(d)** The flow Sankey diagram after the re-planning of the STHK algorithm. **(e)** The trend of the Gini coefficient of taxis distributed in different blocks each day, where the abscissa represents time, and the ordinate is the value of the Gini coefficient. **(f)** The Gini coefficient variation trend of the distribution of taxi transfers in different blocks every day. The actual data is represented by a purple line, and the result of the STHK algorithm is represented by yellow line.

regions, there are heterogeneous features in line with the actual data. For example, comparing Figure 4 (c) and Figure 4 (d), we find that the number of taxis whose destinations are still in the same area accounts for the largest proportion for vehicles departing from one area. In comparison, the number of vehicles from one area to another is quite different. These phenomena show that the STHK algorithm can better retain the characteristics of the actual taxi journeys. Future research could also focus on the correlation between the actual jump distance and the jump intensity. In general, our research confirms that there is also significant heterogeneity in the movement and jump behavior of taxis, similar to other types of past human movement behavior.

The Gini coefficient is a crucial method to describe whether the distribution is uniform. Here we calculate the Gini coefficient of taxi appearance frequency and transfer frequency in different areas every day. Figure 4 (e) reports the Gini coefficient of the frequency of taxi appearances in each area every day. The average Gini coefficient of the actual data is 0.62, with a standard deviation of 0.02, while the average Gini coefficient after STHK algorithm reprogramming is 0.59, with a standard deviation of 0.01. Similarly, the Gini coefficient of the transfer frequency distribution between different regions every day is shown in Figure 4 (f). The average Gini coefficient of the actual data is 0.56, and the standard deviation is 0.01. The average Gini coefficient after the STHK algorithm reprogrammed is 0.68, and the standard deviation is 0.01. These results confirm the heterogeneous characteristics of taxi movement behavior in Figure 4 (a-d). It can be seen that there is a significant gap in the distribution of vehicles and the flow of vehicles in various regions. This gap is closely related to the differences in traffic and population flow in various regions. The STHK algorithm indicates this regular difference, which is consistent with the actual situation, which means that our algorithm could well meet human travel behavior's inherent heterogeneity.

4. Conclusion & Discussion

With the development of information technology, urban travel planning has attracted more and more attention from academia and industry. Recently, research on the minimum fleet size required for urban travel has received increasing attention. Existing research results show that after re-planning travel routes, the size of the fleet can be greatly reduced [22]. However, the above research did not consider how to reduce travel distance and travel time. In order to solve the above problems, we propose an algorithm STHK that comprehensively considers travel distance and time. In this algorithm, we not only consider the driver's endurance time, but more importantly, in the case of conflicts in the algorithm path selection, we consider the time and distance cost of the selected trip, and re-select the path in the global scope. The research results show that the STHK algorithm can not only reduce the size of the fleet required for urban travel, but also greatly reduce the empty time and distance spent in cities itineraries. Moreover, these results are not only applicable to taxis, but also to the currently popular Uber, lyft, Didi and others.

Particularly, we observe that the total off-load distance has dropped sharply from 21,303,835 kilometers to about 9,043,792 kilometers, a drop of 58%. Meanwhile, the total off-load time after the re-planning of the STHK algorithm has been significantly reduced, from 4,609,332 hours to 863,222 hours, a drop of up to 81%. In addition, the daily off-load distance decreased from an average of 760,851 kilometers to 322,993 kilometers, and the daily off-load time decreased from an average of 164,619 hours to an average of 30,829 hours. The average empty distance for each taxi to find the next passenger has dropped from 3,312 meters to 1,393 meters, and the average time has been reduced from 43 minutes to about 8 minutes. More importantly, the itinerary re-planned by the STHK algorithm does maintain the heterogeneous characteristics of human behavior [37, 38, 39, 1, 26, 27].

On the one hand, the above research results show that the urban travel planned by the STHK algorithm can not only greatly reduce the empty distance, but also greatly reduce the empty time, which has the feat of alleviating urban traffic congestion, reducing fossil fuel consumption, narrowing carbon dioxide and other harmful gas emission, and protecting the environment. On the other hand, maintaining the heterogeneous characteristics of human travel behavior also implies that the results of the algorithm can match the travel habits of passengers, being simpler to implement.

Of course, our algorithm only considers the Euclidean distance between two locations when considering the distance, and does not consider the true path distance. In addition, our algorithm is for global static travel planning. Future research should focus more on real-time and local dynamic planning methods. Additionally, another problem is that the current taxi research only focuses on the situation where there is only one start and end point for one journey, and future research should focus on scenarios with multiple start points and end points. Existing studies have shown that there are obvious differences in the travel mode of taxis from the two different perspectives, those of drivers and passengers [26]. Another work shows that due to urban structure and regional restrictions, taxi travel also has a greater impact [27]. Future research should combine the actual characteristics of taxi travel.

References

[1] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, M. Tomasini, Human mobility: Models and applications, Phys. Rep. 734 (2018) 1–74.

[2] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Mod. Phys. 87 (3) (2015) 925–979.

[3] M. Karsai, H.-H. Jo, K. Kaski, Bursty human dynamics, Springer, 2018.

[4] M. Saberi, H. Hamedmoghadam, M. Ashfaq, S. A. Hosseini, Z. Gu, S. Shafiei, D. J. Nair, V. Dixit, L. Gardner, S. T. Waller, M. C. González, A simple contagion process describes spreading of traffic jams in urban networks, Nat. Commun. 11 (1) (2020) 1616.

[5] G. Fontaras, N.-G. Zacharof, B. Ciuffo, Fuel consumption and co2 emissions from passenger cars in europe – laboratory versus real-world emissions, Prog. Energ. Combust. 60 (2017) 97–131.

[6] Y. Wu, S. Zhang, J. Hao, H. Liu, X. Wu, J. Hu, M. P. Walsh, T. J. Wallington, K. M. Zhang, S. Stevanovic, On-road vehicle emissions and their control in china: A review and outlook, Science of The Total Environment 574 (2017) 332–349.

[7] M. Barthelemy, The statistical physics of cities, Nat. Rev. Phys. 1 (6) (2019) 406–415.

[8] P. Zhao, M.-P. Kwan, K. Qin, Uncovering the spatiotemporal patterns of co2 emissions by taxis based on individuals' daily travel, J. Transp. Geogr. 62 (2017) 122–135.

[9] J. Feng, Y. Zhang, W. Song, W. Deng, M. Zhu, Z. Fang, Y. Ye, H. Fang, Z. Wu, S. Lowther, K. C. Jones, X. Wang, Emissions of nitrogen oxides

and volatile organic compounds from liquefied petroleum gas-fueled taxis under idle and cruising modes, Environm. Poll. 267 (2020) 115623.

[10] S. Shaheen, Shared Mobility: The Potential of Ridehailing and Pooling, Island Press/Center for Resource Economics, Washington, DC, 2018, pp. 55–76.

[11] N. Molkenthin, M. Schröder, M. Timme, Scaling laws of collective ride-sharing dynamics, Phys. Rev. Lett. 125 (24) (2020) 248302.

[12] J. B. Greenblatt, S. Saxena, Autonomous taxis could greatly reduce greenhouse-gas emissions of us light-duty vehicles, Nat. Clim. Change 5 (9) (2015) 860–863.

[13] M. Miotti, G. J. Supran, E. J. Kim, J. E. Trancik, Personal vehicles evaluated against climate change mitigation targets, Environ. Sci. Technol. 50 (20) (2016) 10795–10804.

[14] S. Çolak, A. Lima, M. C. González, Understanding congested travel in urban areas, Nat. Commun. 7 (1) (2016) 10793.

[15] P. M. Boesch, F. Ciari, K. W. Axhausen, Autonomous vehicle fleet sizes required to serve different levels of demand, Transport. Res. Rec. 2542 (1) (2016) 111–119.

[16] D.-M. Storch, M. Timme, M. Schröder, Incentive-driven transition to high ride-sharing adoption, Nat. Commun. 12 (1) (2021) 3003.

[17] M. Schröder, D.-M. Storch, P. Marszal, M. Timme, Anomalous supply shortages from dynamic pricing in on-demand mobility, Nat. Commun. 11 (1) (2020) 4831.

[18] G. Berbeglia, J.-F. Cordeau, G. Laporte, Dynamic pickup and delivery problems, Eur. J. Oper. Res. 202 (1) (2010) 8–15.

[19] B. M. Baker, M. A. Ayechew, A genetic algorithm for the vehicle routing problem, Comput. Oper. Res. 30 (5) (2003) 787–800.

[20] J. Yang, P. Jaillet, H. Mahmassani, Real-time multivehicle truckload pickup and delivery problems, Transport. Sci. 38 (2) (2004) 135–148.

[21] J. Alonso-Mora, S. Samaranayake, A. Wallar, E. Frazzoli, D. Rus, On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment, Proc. Nat. Acad. Sci. U.S.A. 114 (3) (2017) 462–467.

[22] M. M. Vazifeh, P. Santi, G. Resta, S. H. Strogatz, C. Ratti, Addressing the minimum fleet problem in on-demand urban mobility, Nature 557 (7706) (2018) 534–538.

[23] J. Hopcroft, R. Karp, An n5/2 algorithm for maximum matchings in bipartite graphs, SIAM J. Comput. 2 (1973) 225–231.

[24] D. J. Fagnant, K. M. Kockelman, Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in austin, texas, Transportation 45 (1) (2018) 143–158.

[25] J. Ke, H. Zheng, H. Yang, X. Chen, Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach, Transport. Res. C-Emer. 85 (2017) 591–608.

[26] W. P. Nie, Z. D. Zhao, S. M. Cai, T. Zhou, Simulating two-phase taxi service process by random walk theory, Chaos 30 (12) (2020) 123121.

[27] W.-P. Nie, Z.-D. Zhao, S.-M. Cai, T. Zhou, Understanding the urban mobility community by taxi travel trajectory, Commun. Nonlinear Sci. (2021) 105863.

[28] A. V. Karzanov, An exact estimate of an algorithm for finding a maximum flow, applied to the problem “on representatives”, Probl. Cybern. 5 (1973) 66–70.

[29] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Recommender systems, Phys. Rep. 519 (1) (2012) 1–49.

[30] H. W. Kuhn, The hungarian method for the assignment problem, Nav. Res. Log. 2 (1) (1955) 83–97.

[31] J. Munkres, Algorithms for the assignment and transportation problems, J. Soc. Ind. Appl. Math. 5 (1) (1957) 32–38.

[32] C. Gini, On the measure of concentration with special reference to income and statistics, Colorado College Publication, General Series 208 (1936) 73–79.

[33] A. Sen, M. A. Sen, S. Amartya, J. E. Foster, J. E. Foster, On economic inequality, Oxford University Press, 1997.

[34] H. Zhang, C. J. R. Sheppard, T. E. Lipman, S. J. Moura, Joint fleet sizing and charging system planning for autonomous electric vehicles, IEEE T. Intell. Transp. 21 (11) (2020) 4725–4738.

[35] H. Yang, X. Qin, J. Ke, J. Ye, Optimizing matching time interval and matching radius in on-demand ride-sourcing markets, Transport. Res. B-Meth. 131 (2020) 84–105.

[36] Z.-G. Huang, J.-Q. Dong, L. Huang, Y.-C. Lai, Universal flux-fluctuation law in small systems, Scient. Rep. 4.

[37] C. Song, T. Koren, P. Wang, A. L. Barabási, Modelling the scaling properties of human mobility, Nat. Phys. 6 (2010) 818–823.

[38] Z.-D. Zhao, Z. Yang, Z. Zhang, T. Zhou, Z.-G. Huang, Y.-C. Lai, Emergence of scaling in human-interest dynamics, Sci. Rep. 3 (2013) 3472.

[39] X.-Y. Yan, W.-X. Wang, Z.-Y. Gao, Y.-C. Lai, Universal model of individual and population mobility on diverse spatial scales, Nat. Commun. 8 (1) (2017) 1639.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (Grant No.6217024046), the Scientific Research Foundation of Shantou University (Grant No. NTF19015), the 2020 Li Ka Shing Foundation Cross-Disciplinary Research (Grant No. 2020LKSFG09D), the Science Promotion Program of UESTC, China (No. Y03111023901014006) and the Guangdong Basic and Applied Basic Research Foundation (Grant No.2021A1515012294). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Devised the research project: ZDZ; Performed numerical simulations: YW and WPN; Analyzed the results: ZDZ, SMC and CG; Wrote the paper: ZDZ, SMC and CG;

Availability of data and material

All data and the relevant codes are available from the corresponding author (zzhidanzhao@gmail.com) on reasonable requests.

Table 1: Basic statistical characteristics of taxi data

Date	**# of Records**	**# of Itineraries**	**# of Taxis**
Aug. 1	55,980,722	228,551	13,244
Aug. 2	56,098,807	229,335	13,275
Aug. 3	55,805,598	234,755	13,178
Aug. 4	56,014,007	234,672	13,368
Aug. 5	56,211,481	236,069	13,219
Aug. 6	55,515,611	229,601	13,151
Aug. 8	54,943,356	230,563	13,643
Aug. 9	53,228,728	229,925	13,659
Aug. 10	52,774,204	234,048	13,689
Aug. 11	53,899,068	254,425	13,675
Aug. 12	53,944,409	259,134	13,550
Aug. 14	54,115,297	240,918	13,642
Aug. 15	54,619,707	253,882	13,656
Aug. 16	53,975,820	229,775	13,730
Aug. 17	54,075,095	262,859	13,613
Aug. 18	53,502,810	263,983	13,696
Aug. 19	54,130,314	245,549	13,396
Aug. 20	54,239,145	237,905	13,519
Aug. 21	54,095,585	236,938	13,645
Aug. 22	54,012,483	235,803	13,330
Aug. 23	53,736,387	233,513	13,194
Aug. 24	52,695,248	231,468	13,629
Aug. 25	53,629,467	251,882	13,807
Aug. 26	53,373,453	260,714	13,950
Aug. 27	53,745,205	261,271	13,958
Aug. 28	54,049,165	255,113	13,924
Aug. 29	54,317,107	267,524	13,753
Aug. 30	53,718,591	254,041	13,681

Table 2: Terminology

Notation	**Description**
G	G is a bipartite graph network, in this study, which is used to indicate the endpoints set of the itineraries.
$U(V)$	$U(V)$ demonstrates two sets of vertices from the bipartition of G, in particular, U and V with an equal number of nodes.
M	M means set of edges between nodes in U and nodes in V .
V_f	V_f presents a vertex with no matching edges connected to it.
ALP	An alternating path (ALP) is a path in which the edges belong alternatively to the matching and not matching. All single edges paths are alternating paths.
AMP	An augmenting path (AMP) is an alternating path that starts from and ends on free vertices. All single edge paths that start and end with free vertices are augmenting paths. The edges in the path alternate between being in the matching and not in the matching.
MCM	The maximum cardinality matching MCM is a set of as many edges as possible with the property that no two edges share an endpoint.