CityLines: Designing Hybrid Hub-and-Spoke Transit System with Urban Big Data

Yanhua Li, *Senior Member, IEEE,* Guanxiong Liu, Zhi-Li Zhang, *Fellow, IEEE*, Jun Luo *Member, IEEE*, Fan Zhang, *Member, IEEE*,

Abstract—Rapid urbanization has posed significant burden on urban transportation infrastructures. In today's cities, both private and public transits have clear limitations to fulfill passengers' needs for quality of experience (QoE): Public transits operate along fixed routes with long wait time and total transit time; Private transits, such as taxis, private shuttles and ride-hailing services, provide point-to-point transits with high trip fare. In this paper, we propose *CityLines*, a transformative urban transit system, employing hybrid hub-and-spoke transit model with shared shuttles. Analogous to Airlines services, the proposed CityLines system routes urban trips among spokes through a few hubs or direct paths, with travel time as short as private transits and fare as low as public transits. CityLines allows both point-to-point connection to improve the passenger QoE, and hub-and-spoke connection to reduce the system operation cost. To evaluate the performance of CityLines, we conduct extensive data-driven experiments using one-month real-world trip demand data (from taxis, buses and subway trains) collected from Shenzhen, China. The results demonstrate that CityLines reduces 12.5%-44% average travel time, and aggregates 8.5%-32.6% more trips with ride-sharing over other implementation baselines.

Index Terms—Hub-and-spoke network, urban computing, spatio-temporal data analytics.

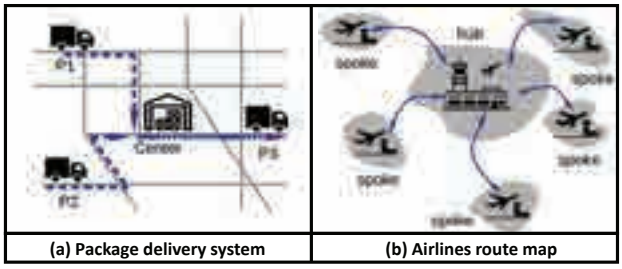
1 Introduction

The past few decades have seen rapid urbanization at the world scale. It is reported that the world urban population has reached 54% in 2014, and it is projected that by 2050, two-thirds of the world population will be urban [2]. The rapid growth in urban population has placed an enormous strain on urban transportation infrastructures. This is particularly the case in developing countries which experience the fastest urbanization, but suffer from far less developed urban transportation infrastructures.

Conventionally, there are two primary models of urban transport systems, namely, *public transit services* such as buses, subway, and *private passenger services* such as taxis, shared shuttles, ride-hailing services (e.g., Uber or Lyft). Both systems have limitations in fulfilling passengers' demands or "quality-of-experience" (QoE), especially during peak demand hours, due to the following fundamental trade-offs in transit service efficiency and costs. Private transits provide exclusive (non-stop) services, thus its transit fare is high, due to the high operation cost. Public transits offer shared rides, thus reducing the cost of operations when there are a significant number of people riding together, say, on a bus. However, existing public transits operate along fixed routes with fixed time tables, where the transit capacity offered do not always match the time-varying trip demands. Consequently, many urban residents rely heavily on private cars and other transport modes (e.g., motor cycles, bikes) to get around a city, creating urban road congestion.

- *Yanhua Li and Guanxiong Liu are with Worcester Polytechnic Institute (WPI), Worcester, MA 01609. E-mail: yli15@wpi.edu. Zhi-Li Zhang is with Computer Science Department at University of Minnesota, Twin Cities, Minneapolis, MN 55455. E-mail: zhzhang@cs.umn.edu. Jun Luo and Fan Zhang are with Shenzhen Institute of Advanced Technologies, Shenzhen China.*

A preliminary version of the results in this paper appeared in [1].



(a) Package delivery system (b) Airlines route map

Fig. 1: Applications of Hub-and-Spoke Model

The aforementioned urban transport systems operate primarily in two modes: *fixed route mode* (with a large number of stops) in public transit services; and *point-to-point mode* in private passenger services. Differing from these two modes, *hub-and-spoke mode*[1] is a system of connections, where all traffic move along spokes connected through a small number of hubs. This mode has been extensively studied in the literature and is commonly used in industry, particularly in Airline route map planning [3], [4], telecommunications [5], freight [6], [7], [8], and package delivery system. Hub-and-spoke mode has advantages over the other two transit modes in the following aspects: It requires less stops/transfers than existing public transits to save on trip time; it requires less routes than private transits, where the smaller number of routes may improve the efficiency of using transportation resources and increase the occupation rate. Figure 1 shows two applications of hub-and-spoke mode in package delivery system and Airlines route map, respectively, where packages and airlines

1. *Fixed route mode* and *hub-and-spoke mode* both allow transfers during a trip, where fixed route mode relies on a large number of densely distributed stops/transfers (e.g., one stop per kilometer) to serve passengers, and hub-and-spoke mode employs very few (usually less than three) hubs per trip to guarantee the quality of experience while aggregating trip demands.

aggregate at and distribute from hubs. Hence, the hub-and-spoke mode offers a great potential to aggregate urban trip demands to leverage economies of scale, while improving users' QoE. However, it can a challenging task to plan and implement the hub-and-spoke mode in urban transportation for the following reasons: *(i)* Urban transits operate on an extremely large spatio-temporal scale, thus it is non-trivial to develop a *scalable* hub-and-spoke network to dynamically serve vast volumes of trip demands over time. In operations research, the *hub location problem (HLP)* has been studied for Industry planning, e.g., airline route planning [3], [4]; these solutions, however, are limited to a maximum scale of 200 regions/spokes. *(ii)* In a real urban area with diverse distributions of trip demands, it is desirable but yet challenging to integrate both point-to-point and hub-and-spoke modes in an adaptive and dynamic fashion.

To tackle these challenges, in this work we propose *CityLines* (in analogy to "Airlines" for flight route services), a scalable dynamic hybrid hub-and-spoke transit system with shared shuttles. The CityLines service relies on a hybrid hub-and-spoke transit network, consisting of a set of inter-connected hub stations in the urban area. A trip demand originated from a small region (referred to as a spoke region) is routed to the destination with a non-stop service (in the point-to-point mode) or via a hub station (in the hub-and-spoke mode). Given a city with n small regions (spokes), if a total budget allows L hubs and M point-to-point transit routes, CityLines aims to find the hub locations and assign urban trip demands to hubs or point-to-point routes, so as to minimize the average travel time. Our main contributions[2] are summarized as follows.

- To scale up the hybrid hub-and-spoke network in CityLines, we propose a two-stage planning framework, including the hub selection stage and the trip assignment stage. The hub selection stage aims to find a small set of high quality candidate regions as hub candidates, so that a maximum number of least travel time paths of trip demands pass through them. Then, the trip assignment stage assigns each trip demand to a hub (for a detour) or a point-to-point transit service, so that the average travel time is minimized.
- To evaluate the performance of our CityLines framework, we conduct experiments on real trajectory data of taxi, bus and subway collected during March 2014 in Shenzhen, China. The results demonstrate that CityLines provides a transformative urban transit service, with travel time as short as private transits and travel cost as low as public transits. Moreover, we deployed a CityLines system [9], and publicized our system code and a part of anonymized urban transit data [10] to allow others to repeat and validate our results, and to (more importantly) facilitate the research in smart transit community.

The remainder of the paper is organized as follows. Section 2 formally defines the problem, presents the overview and outlines the key components of our CityLines framework. Section 3 provides detailed methodology of CityLines framework. Section 4 presents evaluation results over a large-scale urban trip demand data. Related works are discussed in Section 6 and the paper is concluded in Section 7.

2. Note that comparing to the preliminary version of this work in [1], we have (i) introduced a new (optimal hub selection (OHS)) component to significantly promote the system scalability (in Section 3.3); (ii) described the details of our deployed CityLines online system implementation (in Section 5); (iii) presented more comparison results with public and private transit services, and with baselines of hub selection and trip assignments (in Section 4.3 and 4.4.).

2 Overview

In this section, we will motivate and define hybrid hub-and-spoke planning problem, detail the datasets we use, and outline CityLines system framework.

2.1 System Design Trade-offs and Motivations

The choice of urban transit services from a passenger depends on the QoE and cost of the trip, where the QoE hinges on many potential factors, including in-vehicle time, level of inconvenience, etc [11], and the trip cost depends on the service operation cost. Private transit services in general offer high QoE, with low in-vehicle time and high level of convenience, but at a high cost of trip fare. On the other hand, by reducing the operation cost with ride-sharing, public transit services have a lower trip fare, but longer in-vehicle time. Hence, due to the fundamental trade-off between passengers' QoE and operation cost, private and public transit services are operated to meet one of the two aspects, respectively. The next question is *how we can develop a transit service to dynamically serve urban trip demands with travel time as short as taking private transits and trip fare as low as taking public transits?* In this paper, by utilizing the historical trip data from urban transportation systems in Shenzhen, we make the first attempt to develop CityLines, a hybrid hub-and-spoke transit model, that allows an integration of both hub-and-spoke mode (to aggregate trip demands with small number of hubs, thus reduce the operation cost) and point-to-point mode (to reduce the overall trip time, thus to maintain a high passengers' QoE).

2.2 Problem Definition

Thanks to the fast development of location sensing technologies, the increasing prevalence of sensors, mobile devices, and Automated Fare Collection (AFC) devices has led to an explosive increase of the scale of spatio-temporal data, including passenger *trip demands* as defined as follows.

Definition 1 (Trip demand)**.** *A trip demand of a passenger indicates the intent of a passenger to travel from a source location src to a destination location dst from a given starting time t, which can be represented as a triple $\langle src, dst, t\rangle$.*

Passenger trip demands can be obtained from various data sources. For example, the transaction data from AFC devices in buses and subway systems record passenger trip demands at the level of bus stops and subway stations. Taxi GPS trajectory data with occupation information include the trip demands for taxi trips. For urban trip demands, we consider two types of transit modes below, i.e., point-to-point mode and hub-and-spoke mode.

Definition 2 (Point-to-point mode)**.** *With point-to-point mode, a trip demand is served through a direct (usually the shortest or least-cost) path from the source src to the destination dst.*

The urban area consists of small regions, where a trip demand may originate from or destine to. Each of such small regions is referred to as a *spoke*. Some regions, referred to as *hubs*, are deployed with transfer stations, that allow trips to detour at. Given all spoke and hub regions, a hub-and-spoke transit mode can be interpreted as follows.

Definition 3 (hub-and-spoke mode)**.** *With hub-and-spoke mode, a trip demand $\langle src, dst, t\rangle$ is detoured through a small number of ℓ hubs, $h_1, \cdots, h_\ell$ (with $\ell \leq 3$ in general). Thus, the path taken for*

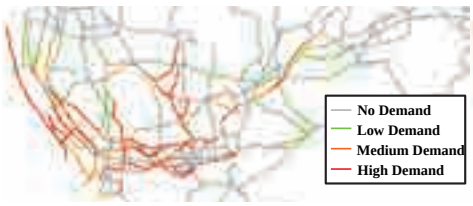


Fig. 2: Trip source locations

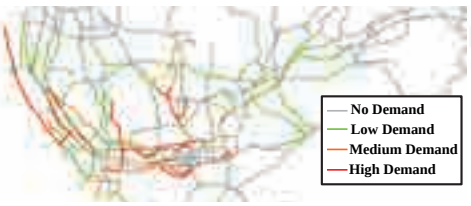


Fig. 3: Trip destination locations

Fig. 4: Shenzhen road map

the trip is $\{src, h_1, \cdots, h_\ell, dst\}$*, and each segment of the path is in general a direct (least-cost) transit.*

Note that the more hubs a trip demand takes, the lower QoE a passenger would receive. In Airlines route planning, one hub detour is commonly used for trip demands. In this paper, to guarantee a high QoE, we allow $\ell = 1$ hub for a trip demand, where our framework also works for cases with $\ell > 1$.

Ideally, for those source-destination location pairs with a large number of trip demands, e.g., commute trips between a residential area and a commercial/working area, point-to-point mode is preferred. On the other hand, for those source destination pairs with less trip demands, hub-and-spoke mode is more promising to aggregate trip demands and reduce the operation cost by leveraging economics of scale. To balance such trade-offs, we propose to investigate the hybrid hub-and-spoke planning problem.

Problem definition. Given a set of n spokes (regions) in an urban area, a set of K trip demands, and a budget of M point-to-point transit routes and L hub stations to deploy, we aim to find the optimal L regions to deploy hub stations and optimal assignment of trip demands to either point-to-point transit or a hub to detour from, so that the average travel time of all trip demands is minimized.

System dynamics. Note that the trip demand distribution changes dramatically over time and follows a stable diurnal pattern. To better cope with the trip demand dynamics, we divide each day into fixed time intervals, and develop CityLines solutions for different intervals. For the rest part of this paper, we focus on solving the hybrid hub-and-spoke planning problem for a given time interval.

2.3 Data Description

To tackle the problem defined above, two real datasets are employed, including (1) trip demand data; (2) road map data. For consistency, all datasets are collected from the same time interval in Shenzhen, China. Below, we describe each of these datasets in details.

Trip demands data are extracted from large GPS trajectory dataset (from taxis) and AFC billing dataset (from buses and subway trains) collected from Shenzhen, China during March 2014. For trip demands from buses and subway trains, we extract their starting and ending stations from the AFC billing data as source and destination locations. On the other hand, we employ taxi GPS data to extract trip demands served by taxis. Each GPS record contains a unique ID, time stamp, latitude, longitude, and passenger indicator. The passenger indicator field is a binary value for taxi data, indicating if a passenger is aboard or not. Hence, a sequence of taxi GPS points with passenger indicator as 1 represent a taxi trip, and the first and last GPS points of the sequence are the source and destination locations (i.e., src and dst) of a trip demand. The time stamp of the starting GPS point is the trip starting time t. Figure 2 and Figure 3 show the geo-distributions of source and destination locations in Shenzhen during the morning rush hours 7–10AM. Note that in the figures, we characterize the trip demands by the density of the events (i.e., the number of source (or destination) locations per hour within a geographic region of 1 km^2 along the road networks). When the density is larger than or equal to 100, the region is considered as a high demand region. A low demand region has the density low than 10 events per hour.

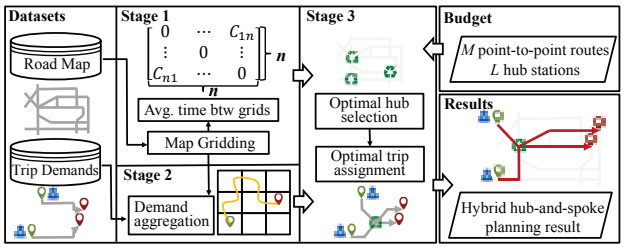


Fig. 5: CityLines Framework

Type	Counts	Type	Counts
Motorway	563	Secondary	868
Trunk	258	Tertiary	1,393
Primary	745	Unclassified	16,829

TABLE 1: Road Map Data in Shenzhen

Road map data. In our study, we use the Google GeoCoding [12] to retrieve the bounding box of Shenzhen. The bounding box is defined between $22.45°$ to $22.70°$ in latitude and $113.75°$ to $114.30°$ in longitude. The covered area is about $1,300 km^2$. Within such a boundary, Shenzhen road map data were obtained from OpenStreetMap [13], which are visualized in Figure 4. The road map data contain six levels of road segments in Shenzhen, which are detailed in Table 1.

2.4 Solution Framework

Figure 5 presents our optimal hybrid hub-and-spoke (OHHS) framework for CityLines system. It takes trip demand data and road map data as inputs. The whole framework consists of three stages in Figure 5: (1) map gridding, (2) trip demand aggregation, and (3) optimal hybrid hub-and-spoke (OHHS) planning.

• **Stage 1 (Map gridding):** The road map is divided into equal grids with a side-length of 0.01 degree in latitude and longitude. Then, a filtering process is conducted to eliminate those grids off the road network, so that the remaining n grids are strongly connected by the road map, namely, each grid can reach any other grid through the road map. We refer to those remaining grids as spokes in the urban area. Then, we estimate average travel time between each spoke pair. Thus, an n by n travel time matrix C is

obtained, which contain the least travel time of each pair of spokes in the urban area.

• **Stage 2 (Trip demand aggregation):** In this stage, all sources and destinations of trip demands are aggregated to the spokes extracted in stage 1. Hence, a trip demand $\langle src, dst, t\rangle$ is aggregated as $\langle s, s', t\rangle$, where s and s' are the spokes where source src and destination dst are located at. Then, a spoke level trip demand matrix V is obtained with each entry V_{ij} representing the number of trip demands originating from spoke i and terminating at spoke j.

• **Stage 3 (Optimal hybrid hub-and-spoke (OHHS) planning):** Given a budget of M point-to-point transit paths, and L hub stations to deploy, we propose a two-step optimization framework to tackle the optimal hybrid hub-and-spoke (OHHS) planning problem, including an optimal hub selection (OHS) step and an optimal trip assignment (OTA) step. The OHS problem is formulated as a maximum coverage problem, that selects $M + L$ high quality hub candidates from n spokes. The OTA problem is formulated as a p hub location problem (p-HLP) problem, which optimally assigns the trips to point-to-point transits or one hub to detour, with the goal of minimizing the average travel time per trip.

Table 2 provides notations used throughout the paper.

Notations	Descriptions
$G = \{g_i\},\ 1 \leq i \leq n$	G is the spoke set of the gridded road map and there are in total $n = \|G\|$ spokes.
$C = \{C_{ij}\}$	C_{ij} is average travel time between a spoke pair (g_i, g_j).
$V = \{V_{ij}\}$	V_{ij} is the number of spoke level trip demands.
$K,\ L,\ M$	K is the total number of trip demands; L (resp. M) is the number of hub stations (resp. point-to-point paths) to be deployed.
$H = \{h_m\},\ 1 \leq m \leq M + L$	H is set of selected physical hub candidates.
$x_k \in \{0, 1\}$	x_k indicates if a spoke k is selected as a hub candidate.
$y_{ij} \in \{0, 1\}$	y_{ij} indicates if a trip demand (g_i, g_j) is covered by hub candidates.
$x_{ij}^m \in \{0, 1\}$	x_{ij}^m indicates if a trip demand (g_i, g_j) detours at a hub candidate h_m.
$y_m \in \{0, 1\}$	y_m indicates if a hub candidate h_m is chosen to deploy a hub.

TABLE 2: Notation Table

3 METHODOLOGY

3.1 Stage 1: Map Gridding

The passenger trip demands (i.e., sources and destinations) are geo-graphically and dynamically distributed across urban areas. In the first stage, the entire urban area needs to be partitioned into spokes (i.e., small regions), so that trip demands with the same source and destination spokes are served in the same fashion, e.g., by the same shuttle at the same time. For the ease of implementation in practice, in this paper, we adopt the gridding based method, which simply partitions the map into equal side-length grids [14], [15]. Moreover, the gridding based method allows us to adjust the side-length of grids, to better examine and understand impacts of the spoke size. Hence, in Stage 1, our approach divides the road map into equal-size grids with a pre-defined side-length s in latitude and longitude. Figure 6 shows all grids (i.e., spokes) in the bounding rectangle region of Shenzhen, China, with $s = 0.01°$. Then, we remove the spokes without a road segment, which are usually located in the no-sense areas, such as ocean or mountain. The remaining spoke set is denoted as G with $n = |G|$ spokes, which can be represented as a graph, with spokes as nodes, connected by the urban road network. Figure 6 highlights (in light color) those $n = 1,018$ spokes on the road network of Shenzhen, China.

Fig. 6: Connected spokes in Shenzhen

Average travel time estimation between spoke pairs. Each spoke grid has a center location, which is not necessarily on a road segment. We first map the center location on a nearest road segment in the spoke, and use the mapped location on the road segment to represent the spoke. Then, for each pair of neighboring spokes g_i and g_j, we can calculate average travel time on the road network from the trajectory data of taxis and buses, denoted as T_{ij}. The matrix $T = [T_{ij}]$ thus represent the adjacency travel time matrix between neighboring spokes. Since the urban road network is well connected, such a spoke graph is strongly connected [16], which means that each spoke g_i has a path to any other spoke g_j. Hence, we can apply the shortest path algorithms, such as Dijkstras and Bellman-Ford algorithms to calculate the least travel time between each spoke pair. We denote the least travel time from spoke g_i to g_j as C_{ij}, and $C = [C_{ij}]$ thus form the least travel time matrix among spokes. The diagonal entries of C indicate the travel time within each spoke. In our study, we set these entries to be 0, namely, CityLines service primarily serves relatively long distance trips. It is more convenient to walk from source to destination for a trip demand within a spoke.

3.2 Stage 2 :Trip Demand Aggregation

Each trip demand $\langle src, dst, t\rangle$ specifies a source location src, and a destination location dst. Given n spokes extracted from stage 1, we now in a position to aggregate all trip demands to spoke pairs, that is, for all trip demands with $src \in g_i$ and $dst \in g_j$, they will be considered in the same group with the source spoke g_i and destination spoke g_j. We denote V_{ij} as the total number of trip demands with source spoke as g_i and destination spoke as g_j. Clearly, $V_{ij} = |\{\langle src, dst, t\rangle | src \in g_i, dst \in g_j\}|$. Then, the volume matrix $V = [V_{ij}]$ indicate the number of pairwise trip demands across the spokes. From our dataset collected from Shenzhen, China (as shown in Figure 2 and 3), the trip demands are distributed unevenly across spoke pairs.

3.3 Stage 3: Optimal Hybrid Hub-and-Spoke (OHHS) Planning

Consider a city with a budget of deploying point-to-point transit service for M spoke pairs, and L hubs for trip demands to detour. Given the spoke set G of n connected spokes, least travel time matrix $C = [C_{ij}]$, and volume matrix $V = [V_{ij}]$ as input, the

hybrid hub-and-spoke planning problem aims to identify M spoke pairs to deploy point-to-point transit, L spokes from G to deploy hubs, and assign each of the rest source-destination spoke pairs to a hub for detour, so as to minimize the average travel time for all trip demands. There are primarily two key challenges in solving this problem: (i) The hub candidate set is the entire spoke set G of size n, where $n = 1,018$ in Shenzhen as discussed in the example in Stage 1. Suppose that there are a total of $L = 10$ hubs to be deployed. The search space of all possible 10 hubs is about a size of $\binom{n}{L} = \frac{n!}{L!(n-L)!} = 6.8 \times 10^{23}$, which is in general unsolvable for a combinatorial optimization without an approximation. (ii) The hub location planning problems have been studied extensively in the literature [17], but none of them consider a scenario with both point-to-point and hub-and-spoke transit modes. Hence, how to formulate the combination of these two transit modes in a single framework is challenging. To address the first challenge, we develop a two-step optimization framework, with step 1 (referred to as optimal hub selection (OHS)) to pre-select a small set of "high quality" hub candidates, and step 2 to find the best L hubs from a much smaller searching space. For the second challenge, we introduce a novel notion of *virtual hub* into the traditional hub location problem to characterize those point-to-point transit mode, namely, all trip demands assigned to the virtual hub are chosen for point-to-point transits. Below, we will elaborate on each of these two steps in details.

3.3.1 Optimal Hub Selection (OHS)

The goal of this step is to pre-select a small set of "high quality" hub candidates from the entire spoke set G of size n, so as to reduce the searching space in the next step when finalizing hub locations. In general, if a hub resides on the least travel time path of a trip demand, it generates the least additional cost, when detouring the trip demand to that hub. In this case, we consider that the hub "covers" the particular trip demand. Hence, given all (spoke-level) trip demands, the single hub candidate that resides on (or covers) the most trip demands is the "best" hub candidate. However, when we look for multiple hub candidates, we want a collection of hub candidates that together cover a maximum number of *unique* trip demands, which may not be the hub candidates with top numbers of covered trip demands, since the coverage of different hub candidates may overlap. Given such intuitions, we formulate our optimal hub (candidate) selection (OHS) problem as follows.

Denote a source-destination spoke pair from spoke g_i to g_j as (g_i, g_j). Given a hub candidate g_k, we denote $S(g_k)$ as the set of source-destination spoke pairs with their least travel time paths going through g_k. Let $\vec{x} = [x_k]$ be a vector of binary hub selection variables, indicating if a spoke $g_k \in G$ is selected as a hub candidate (with $x_k = 1$) or not (with $x_k = 0$). Moreover, we denote $\vec{y} = [y_{ij}]$ as the matrix of binary variables, with y_{ij} indicating if a source-destination spoke pair (g_i, g_j) is covered by the selected candidate hubs (with $y_{ij} = 1$) or not (with $y_{ij} = 0$). We aim to resolve $\vec{x}$, indicating the best hub candidates, and $\vec{y}$, the source-destination spoke pairs covered by the hub candidate set $\vec{x}$, such that the total number of unique trip demands from the covered source-destination spoke pairs is maximized. OHS problem is formally summarized below.

$$\text{max:} \quad \sum_{g_i \in G} \sum_{g_j \in G} V_{ij} y_{ij} \tag{1}$$

$$s.t.: \quad \sum_{g_k \in G} x_k \leq M + L \tag{2}$$

$$\sum_{(g_i, g_j) \in S(g_k)} x_k \geq y_{ij} \qquad \forall g_i, g_j, g_k \in G \tag{3}$$

$$y_{ij}, x_k \in \{0, 1\} \qquad \forall g_i, g_j, g_k \in G \tag{4}$$

The objective function in eq.(1) captures the total number of trip demands being covered by the selected hub candidates. The first constraint (in eq.(2)) indicates that the total number of selected hub candidates is no more than $M + L$, with L as the maximum number of hubs to be deployed, and M as the maximum number of source-destination spoke pairs to be served by point-to-point transit mode. Since the trip demands being covered in step 1 may be served by point-to-point transit mode, selecting $M + L$ hub candidates in step 1 guarantees that we have enough high quality hub candidates for step 2. The second constraint (in eq.(3)) guarantees that if a spoke pair (g_i, g_j) is covered (with $y_{ij} = 1$), at least one spoke g_k, that "covers" (g_i, g_j) should be selected as a hub candidate (i.e., $x_k = 1$). The last constraint (in eq.(4)) specifies that each x_k and y_{ij} is a binary variable.

Our optimal hub selection (OHS) problem is fundamentally a (weighted) maximum coverage problem [18]: Given a number ℓ and n sets of elements, which may have some common elements, we select ℓ of these sets so that the maximum number of unique elements are covered. OHS problem is NP-hard, and there is no polynomial-time algorithm that guarantees to find the optimal solution for all instances unless $P = NP$.

In the literature, there have been a variety of efficient approximation algorithms for solving weighted maximum coverage problem. The generalized maximum coverage algorithm [18] achieves an approximation ratio of $1 - \frac{1}{e} - o(1)$. Moreover, a greedy algorithm for weighted maximum coverage problem has an approximation ratio of $1 - \frac{1}{e}$ [19], [20]. We employ the approximation algorithm in [20] for solving our OHS problem.

3.3.2 Optimal Trip Assignment (OTA)

The output hub candidates from step 1 has significantly reduced the hub selection space from $n = |G|$ to $M + L$. The next step is to further select L hubs from the $M + L$ candidates $\{h_1, \cdots, h_{M+L}\}$, and assign them to spoke pairs, and choose M spoke pairs for point-to-point transit mode, so that the overall average travel time of trip demands is minimized. Without the point-to-point mode part, this problem is a well-studied combinatorial optimization problem, so called, p-HLP (p hub location problem), that aims to select a total of p hubs and assign each trip demand to one and only one hub, to minimize the average trip time. To include the point-to-point transit mode, we introduce a novel notion of *virtual hub*, denoted as h_0, which is not physically one entry from $M+L$ hub candidates. Figure 7 illustrates how the virtual hub h_0 works. All trip demands assigned to h_0 are served by point-to-point transit mode. Instead, a trip demand assigned to a physical hub h_i ($1 \leq i \leq M + L$) will be detoured through h_i during the trip. By introducing the virtual hub h_0, the optimal trip assignment (OTA) problem can be formulated as follow.

Let C_{ij}^k be the travel time for a trip demand from spoke g_i to g_j detoured at hub h_k. Recall that the least travel time from spoke g_i to g_j is C_{ij}. Thus, with a physical hub h_k, we have $C_{ij}^k = C_{ik} +$

C_{kj}; and for the virtual hub h_0, we have $C^0_{ij} = C_{ij}$, since a trip demand assigned to virtual hub h_0 is served with point-to-point transit mode. Let x^k_{ij} be a binary assignment variable indicating if trip demands with source-destination spoke pair (g_i, g_j) are assigned to hub h_k ($x^k_{ij} = 1$) or not ($x^k_{ij} = 0$). Moreover, we denote y_m (with $1 \le m \le M + L$) as a binary selection variable, indicating if a physical hub h_m is selected ($y_m = 1$) or not ($y_m = 0$). We want to resolve y_m, indicating the finally selected L hubs, and x^k_{ij}, the trip assignment to hubs, such that the average travel time of trip demands is minimized. This OTA problem is presented below.

$$\text{min:} \quad \frac{1}{V} \sum_{g_i \in G} \sum_{g_j \in G} \sum_{0 \le k \le M+L} V_{ij} C^k_{ij} x^k_{ij}, \tag{5}$$

$$s.t.: \quad \sum_{0 \le k \le M+L} x^k_{ij} = 1, \qquad \forall g_i, g_j \in G, \tag{6}$$

$$\sum_{g_i \in G} \sum_{g_j \in G} x^0_{ij} \le M, \tag{7}$$

$$\sum_{g_i \in G} \sum_{g_j \in G} V_{ij} x^k_{ij} \le F_k, \qquad 1 \le k \le M + L, \tag{8}$$

$$\sum_{1 \le m \le M+L} y_m \le L, \tag{9}$$

$$y_m \ge x^m_{ij}, \qquad \forall g_i, g_j \in G, 1 \le m \le M + L, \tag{10}$$

$$x^k_{ij} \in \{0, 1\}, \forall g_i, g_j \in G, 0 \le k \le M + L. \tag{11}$$

$$y_m \in \{0, 1\}, 1 \le m \le M + L. \tag{12}$$

The objective function in eq.(5) indicates the average travel time of all trip demands, with $V = \sum_{g_i, g_j \in G} V_{ij}$ as the total number of trip demands to be planned. The constraint in eq.(6) states that each source-destination spoke pair should be served, i.e., by one and only one hub (including the virtual hub). The constraint in eq.(7) ensures that up to M source-destination pairs are served by point-to-point transit mode with direct paths. The constraint in eq.(8) specifies the capacity of each physical hub h_k, namely, the total number of trips going through a hub h_k cannot exceed the hub capacity F_k. The constraint in eq.(9) guarantees that the total number of physical hubs deployed is no more than L. Eq.(10) specifies a validity constraint, where a spoke pair (g_i, g_j) is assigned to a hub candidate h_m, if and only if h_m is selected to deploy a hub, namely, $y_m = 1$. The constraint eq.(11) and eq.(12) indicate that x^k_{ij} and y_m are binary variables.

By introducing the virtual hub h_0 into the formulation, our optimal trip assignment (OTA) problem allows both hub-and-spoke and point-to-point modes. The nice property of OTA formulation is that it still follows p-HLP (p hub location problem). Moreover, with the optimal hub selection step, the searching space for hubs has been reduced from all spokes in G to only $M + L$ hub candidates. In the literature, p-HLP has been extensively studied, with several efficient approximation approaches developed. For examples, Ernst and Krishnamoorthy introduced a 3-index formulation for p-HLP, which enables an LP relaxation based approximation solution [21]. Marin, Canovas and Landete introduced new formulations for p-HLP problem that generalized basic models with providing tighter LP bounds [22]. In this work, we adopt the solution proposed in [21] to solve our OTA problem.

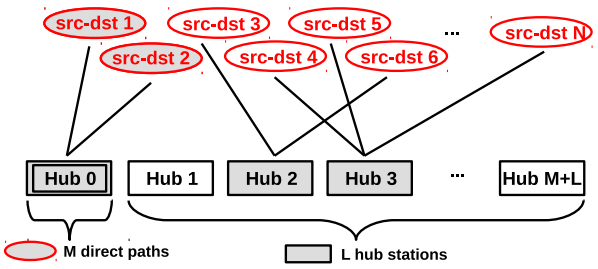


Fig. 7: Illustration of the virtual hub

4 Evaluation

To evaluate the performances of our CityLines system, we conduct comprehensive data-driven experiments using large-scale urban trip demand datasets collected from Shenzhen, China. First of all, the comparison results of CityLines with traditional private and public models clearly demonstrate our advantages in reducing operation cost (i.e., the number of passengers per trip segment) and improving passenger QoE (i.e., average travel time per trip). Secondly, by comparing with baseline algorithms in implementing hybrid hub-and-spoke transit planning, experimental results demonstrate that our CityLines system outperforms all other baselines (i) with 12.5%-44% reduction on average travel time per trip demand, and (ii) with 8.5%-32.6% more aggregated trips via ride-sharing. Below, we elaborate on baseline methods, experiment settings and results.

4.1 Baseline Methods

We will conduct two sets of experiments to *(i)* compare public and private transit models with hybrid hub-and-spoke model employed in CityLines system, *(ii)* compare our proposed optimal hybrid hub-and-spoke (OHHS), i.e., a two-step optimization framework, with other baseline algorithms.

Baseline transit models: We compare private and public transit models with our hybrid hub-and-spoke model.

(1) Private transit model: This model serves trip demands via direct least travel time paths with non-stop service.

(2) Public transit model: This model employs the existing public transit infrastructure (i.e., bus routes and subway lines), to serve all trip demands.

Baselines for hub candidate selection: We compare our optimal hub selection (OHS) method with the two baseline methods below.

(1) Random Selection (RS): This baseline method uniformly at random chooses $M + L$ spokes from G as hub candidates.

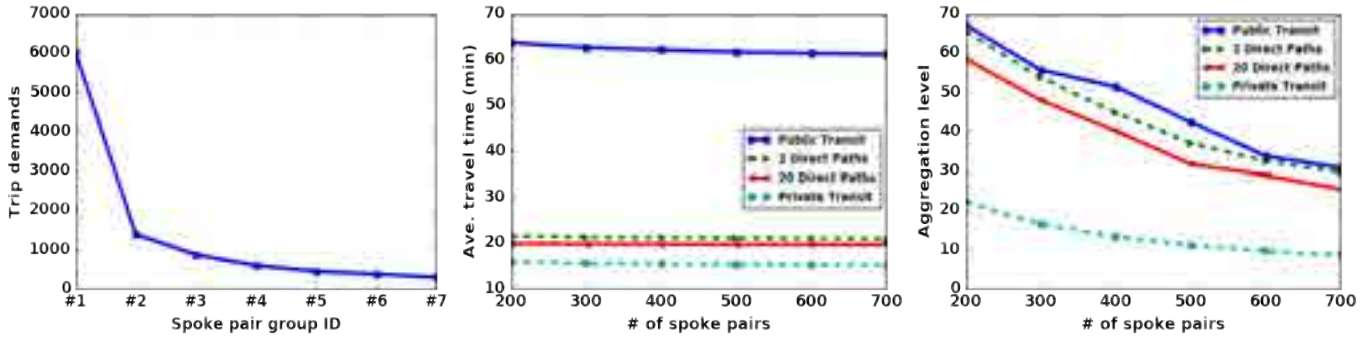
(2) Top Selection (TS): This baseline method selects $M + L$ hub candidates from G with the top numbers of source-destination spoke pairs covered.

Baselines for trip assignment: We compare our optimal trip assignment (OTA) method with the two baseline methods below.

(1) Random Assignment (RA): This baseline method first randomly picks out L hubs from $M+L$ hub candidates, and randomly assigns the trip demands to point-to-point mode or one of hub candidates.

(2) Average Assignment (AA): This baseline method assigns the trip demands to point-to-point mode or one of hub stations, so that each hub (roughly) serves an equal amount of trip demands.

In our experiments, we run the random selection (RS) and random assignment (RA) methods for 50 times and calculate the average results, so as to remove the potential impact of randomness.



(a) Trip demand distribution (b) Average travel time (c) Trip aggregation level

Fig. 8: Comparison of transit models

4.2 Experiment Settings

From the trip demand aggregation stage, we obtain in total $19,428,453$ urban trip demands from taxis, buses and subway in Shenzhen, China, during March 2014. One interesting phenomena we observe from the data is that the trip demand distribution changes dramatically over different time intervals in a day. However, for the same time interval, it stays relatively unchanged over days. This is reasonable since daily urban commute/travel patterns are relatively stable. Hence, to better cope with the dynamics of trip demands, we divide each day into 5 time intervals: 6–11am, 11am–4pm, 4-8pm, 8–12am, 12–6am, develop and apply different hybrid hub-and-spoke plans to each interval. We apply cross-validation mechanism to evaluate our CityLines system: We use a sliding time window of four days. We employ the trip demands of day 1–3 as the input data, and develop the hybrid hub-and-spoke solution. Then, we test the performance of the solution using the trip demand data from day 4. We move the sliding window over the working days in our data, and calculate the average performances for all sliding windows. In this section, we will use the time interval 6–11am, as an example to demonstrate the effectiveness and efficiency of our CityLines system. Results for other time intervals are similar, and are omitted for brevity. Taking the trip demand data during 6–11am on March 12, 2014 as an example, there were in total $202,315$ trip demands in the city. Given those $1,018$ connected spokes obtained, most (more than 90%) of trip demands aggregate to 700 source-destination spoke pairs. We sort all these spoke pairs by their numbers of trip demands in a decreasing order, and divide them into 7 groups, each with 100 spoke pairs. The resulting spoke pair groups with ID $\{\#1, \cdots, \#7\}$ are thus in a descending order in their numbers of trip demands per spoke pair (See Figure 8(a)). We will gradually add trip demands from each group (i.e., high volume group first) into experiments, to evaluate how the problem scale affects the system performance. Table 3 lists configurations used in our evaluation.

For different planning methods, we evaluate operation cost using the trip aggregation level, and evaluate the passenger QoE using average travel time. Moreover, we use the number of covered unique trips to evaluate the quality of hub candidates selected in the optimal hub selection (OHS) step. These metrics are detailed below.

Average travel time. Given a path planned for a trip demand $tr = \langle src, dst, t\rangle$ from the source to the destination, i.e., $\{g_1, \cdots, g_\ell\}$, the total travel time is given by $\sum_{2\le i\le \ell} T_{i-1,i}$. The average travel time of all trip demands characterizes the quality of experience passengers receive from the planning strategy. The lower the time is, the higher QoE passengers experience.

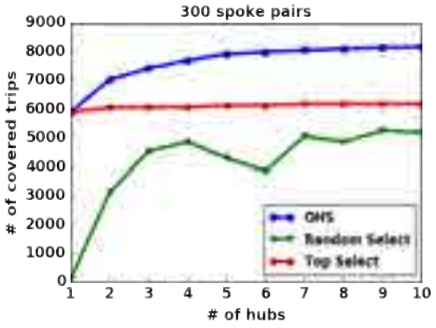
Trip aggregation level (of trip demands). Given a planning method, each trip demand traverses a few *trip segments*. For example, in public transit model, the trips are divided into small trip segments between consecutive stop pairs. In CityLines service, each trip consists of spoke-to-hub and hub-to-spoke trip segments. In private transit model, each spoke pair maintains a unique trip segment as the direct path. Since trip demands may share the trip segments, each trip segment has a certain number of shared trip demands. The average number of shared trip demands per trip segment indicates the ride-sharing level, or trip aggregation level of the planning method. The higher the trip aggregation level is, the lower the operation cost is.

Hub coverage. To reduce the computational cost, we pre-select a small set of "high quality" hub candidates from the spoke set G. Intuitively, the hubs residing on the least travel time paths are with good quality, in terms of generating additional travel time. Hence, we evaluate the quality of a selected set of hub candidates, using the number of unique least travel time paths they covered, (in short, referred to as hub coverage).

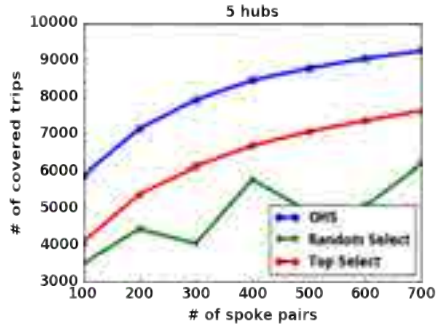
Running time. For the same number of spoke pairs, hubs, and directed paths, we evaluate the computational time (i.e., running time) by comparing our scalable OHHS algorithm to the basic OHHS algorithm proposed in [1] (in short, OHHS-Basic). Figure 12 shows the results with the problem scale ranging from 20 to 700 spoke pairs. The planning budget includes 10 hubs and 5 directed paths. The results clearly indicate that our scalable 2-stage OHHS framework only takes less than 3 minutes for a problem with 700 spoke pairs. On the other hand, when directly

spoke pairs	$\{100, 200, \cdots, 700\}$
# of hubs	$\{1, 2, \cdots, 10\}$
# direct paths	$\{1, 2, \cdots, 10\}$
transit model	hybrid hub-and-spoke, public transit, private transit
hub selection	OHS, Top Selection (TS), Random Selection (RS)
trip assignment	OTA, Average Assign (AA), Random Assign (RA)
hybrid hub-and-spoke planning	OHHS, TS-AA, TS-RA, RS-AA, RS-RA, OHHS-Basic [1]

TABLE 3: Evaluation configurations

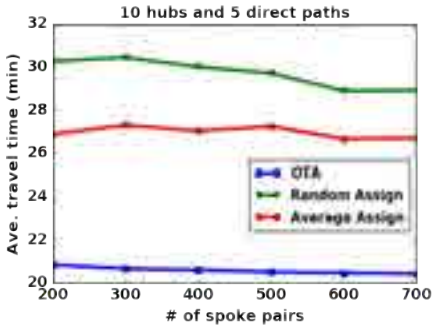


(a) Hub coverage over # hubs

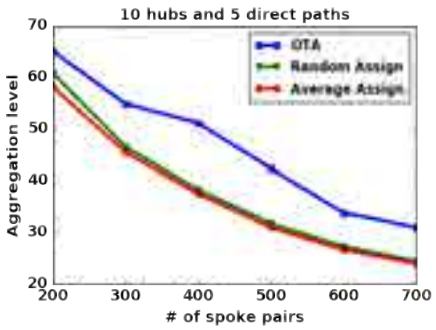


(b) Hub coverage over spoke pairs

Fig. 9: Hub candidate selection

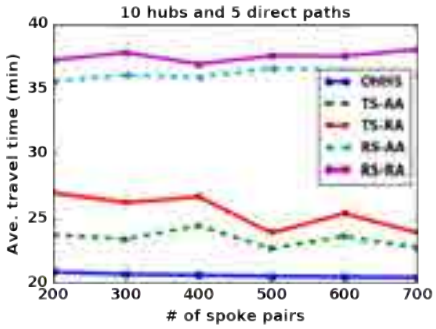


(a) Average travel time over spoke pairs

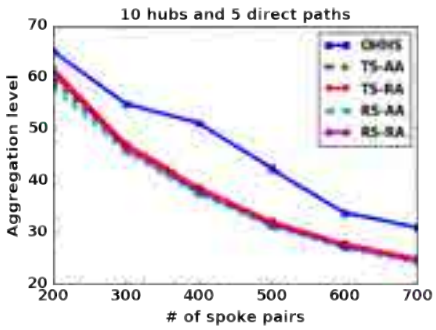


(b) Trip aggregation level over spoke pairs

Fig. 10: Trip assignment



(a) Average travel time over spoke pairs



(b) Trip aggregation level over spoke pairs

Fig. 11: Hybrid framework

solving the hub assignment problem, the running time of OHHS-Basic increases dramatically from 20 seconds (for 20 spoke pairs) to 77 minutes (for 150 spoke pairs). OHHS-Basic fails to find results for a problem with more than 150 spoke pairs due to the exponentially increased computational complexity.

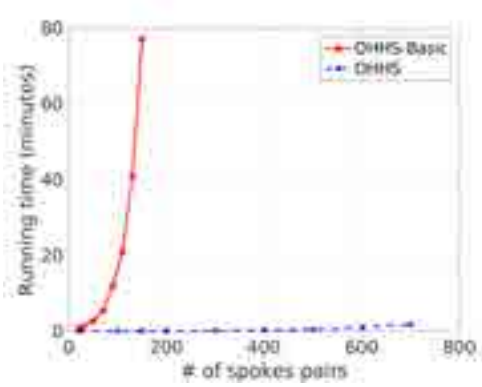


Fig. 12: Running time comparison

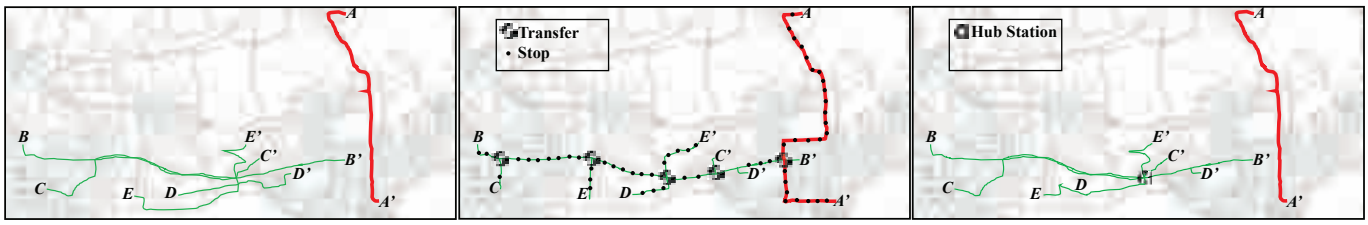
4.3 Comparison of Transit Models

Figure 8(b)–(c) show the comparison between three different transit models, including public transit, private transit, and our hybrid hub-and-spoke models. As more trip demands being included, the results show clearly the trade-off between the three transit models, in terms of the average travel time (as a measure of passenger QoE) and the trip aggregation level (quantifying the operation cost): (i) Private transit model always achieves the lowest average travel time for trip demands, which is reasonable, since the private transit model takes the least travel time paths for trips. However, due to the low ride-sharing rate, the trip aggregation level is always the lowest comparing to other models, thus leads to high operation cost. (ii) On the other hand, by coordinating trip demands at a large number of bus stops and subway stations, public transit model always achieves the highest trip aggregation level than other transit models, thus significantly reduces the operation cost. However, high transition time incurred at stops and stations leads to the highest travel time, over other models. (iii) By allowing both hub-and-spoke and point-to-point connections, our hybrid hub-and-spoke model can dedicate necessary point-to-point resources to high-volume spoke pairs, while aggregating low-volume spoke pairs via hubs. As a result, our hybrid hub-and-spoke model can achieve as low average travel time as private transit model, and as high trip aggregation level as public transit model.

4.4 Hybrid Hub-and-Spoke Planning

Given the clear advantages of our hybrid hub-and-spoke model over the traditional private and public transit models, we now move on to evaluate our CityLines system (as a 2-step optimization solution) by comparing it with baseline implementation algorithms.

Step 1: Hub candidate selection. Figure 9(a)–(b) presents the comparison results on the hub coverage, between our OHS method and two baseline algorithms, including top selection and random selection. As we increase the number of hub candidates, Figure 9(a) shows that hub candidates selected by our OHS method always cover more trip demands than random selection, and top selection methods. when the total number of hub candidates to be selected is small, our OHS methods can select high quality hub candidates that cover up to 12 times more trip demands (about 6000 trip demands), than random selection method (about



(a) Private transit (b) Public transit (c) CityLines service

Fig. 13: Case Studies

500 trip demands). When the number of selected hub candidates is large, e.g., 10 hub candidates, the winning margin of our OHS (with about 8000 trip demands) is about 1.6 times over baseline algorithms (about 5000 trip demands). Consistent results (in Figure 9(b)) are obtained when we increase the number of spoke pairs. With 5 hubs for 300 spoke pairs, our OHS method can select hub candidates that covers twice trip demands (about 7900 trip demands) of random selection method (about 3900 trip demands). When more spoke pairs are included (say, 700 spoke pairs), the hub candidates selected by our OHS method cover 9200 trip demands, which is about 2.1 times hub coverage of random selection method (of 6100 trip demands). Overall, OHS method selects hub candidates with 1.6 to 12 times hub coverage than other baselines.

Step 2: Trip assignment. Figure 10(a)–(b) show the comparison results between our optimal trip assignment (OTA) method with two baselines, including random assignment (RA) and average assignment (AA). To guarantee a fair comparison among different trip assignment methods, we use the same set of hub candidates selected in step 1 by OHS method. Figure 10(a) shows results in average travel time, where our OTA always achieves the lowest average travel time, with an average of 7%-31% reduction than other baselines. On the other hand, Figure 10(b) shows results in trip aggregation level: our OTA always has the highest trip aggregation level. Given 10 hubs and 5 direct paths, our OTA method has around 31–64 trips aggregated per trip segment, while baseline methods only have about 25–59 trips aggregated per trip segment, which leads to a total of 8%-24% improvement in trip aggregation (thus reduction in operation cost).

Hybrid hub-and-spoke planning. Figure 11(a) shows that our OHHS framework always achieves the lowest average travel time with about 21 min, while other baseline methods lead to much higher average travel time ranging from 24 to 38 min. Thus, our framework achieves a total of 12.5% to 44% reduction on average travel time. When measuring the trip aggregation level (Figure 11(b)), our OHHS framework always has the highest number of aggregated trips, with a total of 8.5%–32.6% improvement over baseline algorithms.

4.5 Case Studies

Figure 13(a)–(c) show an example with real trip demands, which demonstrate the effectiveness of CityLines service by comparing it with private and public transit services. We extract a small set of trip demands during 6–11am in March 12, 2014, from Shenzhen, China. The trip demand set includes a total of 1,274 trip demands with 5 source spokes and 5 destination spokes. One source-destination pair (from spoke A to A') is with the highest trip demand volume, i.e., 473 trip demands. Moreover, each source (from B, C, D) has some trip demands (ranging within 58 – 118) to each destination (in B', C', D'), and E has 77 trip demands to E'. Figure 13(a)–(c) show the trip planning solutions using three transit models, including private transit, public transit, and CityLines service (with one hub and one direct path as the budget). Our results show that private transit and CityLines lead to similar average travel time, as 23 and 26 minutes, respectively, and public transit has 47 minutes average travel time due to the large number of stops and transfers during the trips. On the other hand, public transit and CityLines enable similarly high aggregation levels, with 168 and 155 aggregated demands per trip segment, where private transit leads to only 112 aggregated demands[3], due to the distinct least travel time paths employed.

5 System Deployment

In this section, we describe the details of our deployed system.

Fig. 14: System Interface.

Our CityLines system is publicly available online [9], where the website user interface is implemented using bootstrap, Java, OpenStreetMap, and the system is deployed on a WPI server. Figure 14 is an example of the system interface. The system allows users to interact with it using different parameters to obtain hub-and-spoke network recommendations in a real-time fashion. The interface contains the following components:

Parameters. In the system interface (as shown in Figure 14), there are a few parameters that allow users to choose the desired deployment settings, such as the time interval of interests (in each two hours of a day), total number of hubs (L) and number of directed paths (M) to deploy. Once the user defines and chooses those parameters and presses the button "Generate", the planned hub-and-spoke network will be displayed. Moreover, to achieve

3. Note that the aggregation level of private transits is calculated without considering vehicle capacity. When using taxis, the aggregation level is up to 4, i.e., taxi capacity.

better visualizations for the project, we also developed a drop-down menu to the right of the "Generate" button, with which users can choose the background color patterns: *auto-change*, the background color will change based on the time interval a user chooses; *day-time,* bright white background color all the time; *night-time,* dark grey background color all the time.

Result. On the right hand side of the demo page, the planned hub-and-spoke network will be shown, with red dots representing the planned hub station locations, and green dots representing the source and destination spokes to be served by directed paths. The red paths highlighted between the source and destination spokes are the direct path routes.

Data and Code Sharing: We also make our code and (a subset of) data publicly available on the project webpage [10]. We believe that this will not only allow other researchers to repeat and validate our results, but also facilitate the research community.

6 Related Work

To the best of our knowledge, we are the first to investigate hybrid hub-and-spoke transit model in solving urban transit planning problem. We discuss two closely related topics to our work: (1)urban computing and (2) hub-and-spoke network planning.

Urban computing integrates urban sensing, data management and data analytic together as a unified process to explore, analyze and solve existing critical problems in urban area such as traffic congestion, energy consumption and pollution [23]. For example, by analyzing a large-scale real electric taxi trajectory dataset, authors in [15], [24] develop scalable charging station placement strategies to reduce seeking and waiting time for electric vehicles in urban areas. In [25], [26], the authors developed novel models to predict the road traffic and crowd flows in subway stations. However, none of the existing work addresses the fundamental transit planning problem by employing the novel hybrid hub-and-spoke transit model. Our study shed lights on the opportunity of transforming the urban transit model to provide higher quality of services to passengers.

Hub-and-spoke network planning has been extensively studied in the literature, where all trip demands need to be detoured via hubs to their destination spokes [17]. [27], [28] all attempt to address a *single allocation hub-and-spoke problem*, where multiple hubs are deployed, but all trips from the same spoke have to detour at the same hub. [3], [29] develop solutions to *multiple allocation hub-and-spoke problem*, where trips from the same spoke, with different destination can potentially employ different hubs for detour. However, few works have addressed the hybrid hub-and-spoke network planning by allowing both point-to-point and hub-and-spoke services. Moreover, the existing solutions can only solve a hub-and-spoke problem with limited scale, say, 200 spokes and 10 hubs, which is not applicable to large-scale urban trip planning scenarios. Our CityLines system design aims to fundamentally address these two challenges to develop a *scalable* trip planning service with low system operation cost, and high passenger QoE.

7 Conclusion

In this paper, we make the first attempt to develop CityLines system for urban scale transportation services, that employs a hybrid hub-and-spoke transit model. The model allows both point-to-point connection to improve the passenger quality of experience, and hub-and-spoke connection to reduce the system operation cost. CityLines employs a two-step optimization framework to enable a scalable solution to the optimal hybrid hub-and-spoke planning problem. Comparing with other implementation baselines, the evaluation results (obtained with real world transit data) demonstrate that CityLines reduces 12.5%-44% average travel time, and aggregates 8.5%-32.6% more trips with ride-sharing.

8 Acknowledgments

Yanhua Li was supported in part by NSF CRII grant CNS-1657350 and a research grant from DiDi Chuxing Research. Zhi-Li Zhang was supported in part by DTRA grant HDTRA1-14-1-0040, DoD ARO MURI Award W911NF-12-1-0385 and NSF grants CNS-1411636 and CNS-1618339.

References

[1] G. Liu, Y. Li, Z.-L. Zhang, J. Luo, and F. Zhang, "Citylines: Hybrid hub-and-spoke urban transit system," in *SIGSPATIAL GIS*, 2017.

[2] D. of Economic and S. Affairs, "World urbanization prospects 2014," https://esa.un.org/unpd/wup/Publications/Files/WUP2014-Highlights.pdf.

[3] H. Karimi and M. Bashiri, "Hub covering location problems with different coverage types," *Scientia Iranica*, 2011.

[4] A. Menou, A. Benallou, R. Lahdelma, and P. Salminen, "Decision support for centralizing cargo at a moroccan airport hub using stochastic multicriteria acceptability analysis," *European Journal of Operational Research*, 2010.

[5] H. Kim and M. E. O'Kelly, "Reliable p-hub location problems in telecommunication networks," *Geographical Analysis*, 2009.

[6] S. Çetiner, C. Sepil, and H. Süral, "Hubbing and routing in postal delivery systems," *Annals of Operations Research*, 2010.

[7] J. J. Wang and M. C. Cheng, "From a hub port city to a global supply chain management center: a case study of hong kong," *Journal of Transport Geography*, 2010.

[8] R. Ishfaq and C. R. Sox, "Intermodal logistics: The interplay of financial, operational and service issues," *Transportation Research Part E: Logistics and Transportation Review*, 2010.

[9] CityLines Research Group, "Citylines system," https://wpi.edu/~yli15/CityLines/.

[10] ——, "Citylines system data and code," https://wpi.edu/~yli15/CityLines/info.html.

[11] C. Kumar, D. Basu, and B. Maitra, "Modeling generalized cost of travel for rural bus users: a case study," *Journal of Public Transportation*, 2004.

[12] "Google GeoCoding," https://developers.google.com/maps/documentation/geocoding/.

[13] "OpenStreetMap," http://www.openstreetmap.org/.

[14] Y. Li, M. Steiner, J. Bao, L. Wang, and T. Zhu, "Region sampling and estimation of geosocial data with dynamic range calibration," in *ICDE*, 2014.

[15] Y. Li, J. Luo, C.-Y. Chow, K.-L. Chan, Y. Ding, and F. Zhang, "Growing the charging station network for electric vehicles with trajectory data analytics," in *ICDE*, 2015.

[16] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, "Giant strongly connected component of directed networks," *Physical Review E*, 2001.

[17] R. Z. Farahani, M. Hekmatfar, A. B. Arabani, and E. Nikbakhsh, "Hub location problems: A review of models, classification, solution techniques, and applications," *Computers & Industrial Engineering*, 2013.

[18] D. S. Hochbaum, "Approximation algorithms for the set covering and vertex cover problems," *SIAM Journal on computing*, 1982.

[19] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functionsi," *Mathematical Programming*, 1978.

[20] D. S. Hochbaum, "Approximating covering and packing problems: set cover, vertex cover, independent set, and related problems," in *Approximation algorithms for NP-hard problems*. PWS Publishing Co., 1996.

[21] A. T. Ernst and M. Krishnamoorthy, "Exact and heuristic algorithms for the uncapacitated multiple allocation p-hub median problem," *European Journal of Operational Research*, 1998.

[22] A. Marın, L. Cánovas, and M. Landete, "New formulations for the uncapacitated multiple allocation hub location problem," *European Journal of Operational Research*, 2006.

[23] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," *ACM TIST*, 2014.
[24] C. Liu, K. Deng, C. Li, J. Li, Y. Li, and J. Luo, "The optimal distribution of electric-vehicle chargers across a city," in *ICDM*, 2016.
[25] X. Liu, X. Kong, and Y. Li, "Collective traffic prediction with partially observed traffic history using location-base social media," *CIKM*, 2016.
[26] E. Toto, E. A. Rundensteiner, Y. Li, R. Jordan, M. Ishutkina, K. Claypool, J. Luo, and F. Zhang, "Pulse: A real time system for crowd flow prediction at metropolitan subway stations," in *ECML-PKDD*, 2016.
[27] K. Smith, M. Krishnamoorthy, and M. Palaniswami, "Neural versus traditional approaches to the location of interacting hub facilities," *Location Science*, 1996.
[28] J. Kratica, Z. Stanimirović, D. Tošić, and V. Filipović, "Two genetic algorithms for solving the uncapacitated single allocation p-hub median problem," *European Journal of Operational Research*, 2007.
[29] B. Qu and K. Weng, "Path relinking approach for multiple allocation hub maximal covering problem," *Computers & Mathematics with Applications*, 2009.

Yanhua Li (S'09–M'13–SM'16) received two Ph.D. degrees in electrical engineering from Beijing University of Posts and Telecommunications, Beijing in China in 2009 and in computer science from University of Minnesota at Twin Cities in 2013, respectively. He has worked as a researcher in HUAWEI Noah's Ark LAB at Hong Kong from Aug 2013 to Dec 2014, and has interned in Bell Labs in New Jersey, Microsoft Research Asia, and HUAWEI research labs of America from 2011 to 2013. He is currently an Assistant Professor in the Department of Computer Science at Worcester Polytechnic Institute (WPI) in Worcester, MA. His research interests are big data analytics and urban computing in many contexts, including urban network data analytics and management, urban planning and optimization.

Guanxiong Liu received the B.S. and M.S. degrees in electrical engineering from southeast university, Jiangsu, China, in 2013, and Worcester Polytechnic Institute (WPI) in 2015, respectively. He was a PhD student in Data Science at WPI, when he conducted this research.

Zhi-Li Zhang (M'97–SM'11–F'12) received the B.S. degree in computer science from Nanjing University, Jiangsu, China, in 1986, and the M.S. and Ph.D. degrees in computer science from the University of Massachusetts Amherst, Amherst, in 1992 and 1997, respectively. In 1997, he joined the Computer Science and Engineering faculty at the University of Minnesota, Minneapolis, MN, where he is currently a Professor. From 1987 to 1990, he conducted research with the Computer Science Department, Aarhus University, Aarhus, Denmark, under a fellowship from the Chinese National Committee for Education. He has held visiting positions with Sprint Advanced Technology Labs, Burlingame, CA; IBM T. J. Watson Research Center, Yorktown Heights, NY; Fujitsu Labs of America, Sunnyvale, CA; Microsoft Research China, Beijing, China; and INRIA, Sophia-Antipolis, France.

Jun Luo is a principal researcher at Lenovo Machine Intelligence Center in Hong Kong. He received his PhD degree in computer science from the University of Texas at Dallas, USA, in 2006. His research interests include big data, machine learning, spatial temporal data mining and computational geometry. He has published over 90 journal and conference papers in these areas.

Fan Zhang received the B.S. degree in communication engineering and the M.S. and Ph.D. degrees in communication and information system from the Huazhong University of Science and Technology, Wuhan, China, in 2002, 2004, and 2008, respectively. He is currently an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. From 2008 to 2009, he was a Senior Research Associate with the City University of Hong Kong. From 2009 to 2011, he was a Post-Doctoral Fellow with the University of New Mexico and the University of Nebraska-Lincoln, USA. His research topics include cloud computing, big data, cyber-physical systems, privacy, and security in cloud computing.