Received March 31, 2019, accepted June 11, 2019, date of publication June 24, 2019, date of current version July 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2924434

A Probabilistic Approach for Maximizing Travel Journey WiFi Coverage Using Mobile Crowdsourced Services

AHMED BEN SAID AND ABDELKARIM ERRADI
Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar

Corresponding author: Ahmed Ben Said (abensaid@qu.edu.qa)

This work was supported by the Qatar National Research Fund (a member of the Qatar Foundation) through the National Priorities Research Program (NPRP) under Grant NPRP9-224-1-049.

ABSTRACT A public transport journey planning service often yields multiple alternative journeys plans to get from a source to a destination. In addition to journey preferences, such as connecting time and walking distance, passengers can select the optimal plan based on mobile crowdsourced WiFi coverage available along the journey. This requires discovering mobile crowdsourced WiFi services available along the journey path. However, this task is challenging due to the uncertain availability of discovered services. To enhance the availability of WiFi coverage, we propose a probabilistic approach to discover groups of available crowdsourced WiFi services along with the journey segments. We first analyze the log of their trajectories and use a density estimation technique to discover reference spots representing the frequently visited locations. Then, a joint discrete Fourier transform and autocorrelation analysis are applied to mine the periods of the presence of moving crowdsourced services with respect to each reference spot. A low-complexity cluster analysis based on Jensen–Shannon divergence is then used to mine the periodic movement behaviors of services during the identified periods. Finally, mobile crowdsourced WiFi services that are simultaneously available at intersecting reference spots are grouped. The QoS of discovered groups is computed in terms of availability confidence, failover capacity, aggregated bandwidth capacity, and coverage. Additionally, we propose an algorithm to determine the best public transport journey plan offering based on the QoS of available WiFi service groups along the journey path. We conduct a comprehensive comparative study to validate the effectiveness of the proposed framework.

INDEX TERMS Journey planning, crowdsourced WiFi services, reference spots, service failover.

I. INTRODUCTION

The recent technological achievements in mobile and IoT infrastructures have contributed to the development of smarter devices such as smartphones and smart watches. These devices and gadgets are equipped with multiples sensors capable of capturing rich data e.g. position, temperature, pulse. As a result, a new generation of applications has emerged by soliciting the contribution of the crowd. The new paradigm is termed as mobile crowdsourcing [1], [8].

Recently, mobile crowdsourcing has become an attractive research topic. A great body of works makes use of crowd participation to drive new achievements particularly in context of smart cities. In [2], authors used mobile crowdsourcing to track the available parking spots and suggested key guidelines to effectively harness crowd participation. Dong *et al.* [3] proposed a mobile application to track gas price across city by relying on crowd participation. The mobile camera is triggered when the participant is in the vicinity of a gas station. Then, an image processing algorithm extracts the fuel price. In context of urban city management, MySanJose mobile App. allows residents of San Jose to report any issue such as pothole or street light problem. Similarly, NYC311 mobile App. allows residents of New York to report potholes, snowy streets or sidewalks and more. In [4], a feasibility study of mobile crowdsourcing is conducted to manage municipality resources. This study argued that the success of large-scale crowdsourcing solution heavily depends on user preference and behavior of the citizens. Authors conducted a survey on 1300 participants to discover key concerns and behavioral

The associate editor coordinating the review of this manuscript and approving it for publication was Shuiguang Deng.

preference regarding municipality services. Based on the findings, authors proposed urban services reporting platform and emphasized on the importance of matching between the mobility patterns against the location of the tasks.

Public transportation service is a key component of any smart city initiative and project. Classic transportation service relied on static schedule as advertised by the service provider (e.g. the Metropolitan Transportation Authority) to provide commuters with journey plan from source to destination. The recent advances in data analytics have been very useful to further drive transportation service toward better QoS. Pelletier *et al.* [6] reviewed how smart transport card can be useful not only in daily transit system operation but also in long-term strategic planning of the transport network. Zheng *et al.* [5] used taxicabs logs to discover regions with salient traffic problems. The findings allow also to evaluate how effective the undertaking measures by city planners such as new road segments and subway lines. Furthermore, results can be used to recommend changes to future city planning projects. The aforementioned approaches are termed as infrastructure-based since they rely on costly built-in sensors deployed by service providers. Lu *et al.* [7] further argued that infrastructure-based approach does not capture the quality of commuting experience such as how long a person had to wait at a taxi stands for boarding. Authors advocated for the interoperability of infrastructure approach and participatory sensing or mobile crowdsensing to further get better insights about transportation and enhance the QoS. Neiat *et al.* [10] proposed a novel spatiotemporal abstraction on the cloud of the sensors embedded on public transport transit vehicle (tram, bus ...). This abstraction is termed sensor-cloud service. It enables composition between services in order to provide the best QoS journey plan from source to destination using public transport service. This service abstraction is driven by the ease of access, storage and management not to mention the low cost and wide availability of the cloud. Sensors, embedded on the transit vehicle, are made available on the cloud as a service in space and time. Spatiotemporal composition is conducted since one single service usually does not fulfill user journey preference (waiting time, travel time ...). The sensor-cloud service is abstracted in space and time as a line segment from source to destination with functional and non-functional attributes. In [11], authors proposed spatiotemporal model to provide better quality of experience of journey planning service by adding a second layer of composition of crowdsourced WiFi coverage service. Specifically, suppose Sarah has extra data balance till the next billing period, she wants to share WiFi with others in return for a monetary compensation. In this context, the sensor is the smartphone providing WiFi sharing. The crowdsourced WiFi service is abstracted in space and time with functional and non-functional attributes. Given a set of optimal journey plans, the proposed framework identifies the best composition of crowdsourced WiFi services and consequently, the best optimal plan in terms of WiFi coverage. However, the functional and non-functional attributes of the crowdsourced service are considered static and known beforehand. In addition, in the highly dynamic environment of sensed data, non functional attributes of a crowdsourced service are very likely to fluctuate. Indeed, a participant service might not be consistently available or no longer provide a satisfactory QoS. Therefore, a failure mechanism to assure the continuation of the service is mandatory. In [12], authors proposed a novel algorithm based on the well known $D^{*}Lite$ algorithm, termed $STD^{*}Lite$, for journey planning. $D^{*}Lite$ is a dynamic shortest path algorithm with wide usage in robotics and autonomous vehicle navigation. Whenever the QoS is no longer satisfactory or a component crowdsourced service becomes unavailable, $STD^{*}Lite$ is invoked to find an alternative crowdsourced service composition.

In our previous work [9], we focused on undeterministic crowdsourced WiFi services. Undeterministic refers to the lack of a priori knowledge on the availability of a crowdsourced service at a certain location for a certain period. More precisely, We formulated the task as a two-stage learning framework. In the first stage, we sought to find which service(s) is (are) available at the area where a query for WiFi access is received. By leveraging spatiotemporal features of the historical service availability, the problem is formulated as a classification procedure. A Deep Neural Network (DNN) model, once trained, can be queried to predict the set of potential available services. In the second stage, the objective is to determine the duration of the service availability.

In this work, we focus on providing the best journey plan using both the service paradigm and spatiotemporal data analytics. Specifically, given a set of optimal journey plans, the objective is to identify the optimal plan, i.e. the one with the best WiFi coverage provided by the crowd. The service paradigm is used to abstract the crowdsourced service. Furthermore, we assume the realistic scenario in which the availability of crowdsourced WiFi service is unknown. This uncertainty property has motivated us to apply a probabilistic periodic behavior mining approach to determine the availability of a crowdsourced service. This strategy is driven by the availability of spatiotemporal data and the intrinsic periodic property of human behavior. In addition, to maintain the continuity of the WiFi coverage, we consider the grouping pattern of crowdsourced services in order to ensure the availability of failover services in the vicinity.

Our contribution is summarized as follows:

- We propose a framework to determine the optimal plan based on the service paradigm and the periodicity of crowdsourced service providers.
- We propose a novel abstraction of crowdsourced WiFi coverage service based on the grouping pattern. This abstraction is necessary in order to ensure the availability of failover service.
- We introduce a low complexity periodic behavior analysis approach that makes use of spatiotemporal records of service providers. Thus, the periodic behavior of each provider is derived and the probability of its availability at a particular spot and particular time is obtained.

- We propose a simple and effective algorithm to determine the best journey plan in terms of WiFi coverage that takes into account the QoS of both individual and group of crowdsourced services.
- We demonstrate through synthetic data and real-world scenario the effectiveness of our framework in mining the periodic behavior and finding the best journey plan.

II. JOURNEY PLANNING AND CROWDSOURCED WiFi COVERAGE SERVICE

In this section, we start by a motivation scenario and present our journey planning and crowdsourced WiFi coverage service framework.

A. MOTIVATION SCENARIO

John would like to travel from point A to B using public transport service. While commuting, he would like to enjoy internet access. In this scenario, illustrated in Fig. 1, WiFi is provided by the crowd. Indeed, suppose Sarah has extra data balance till the next billing period. We suppose she is well-incentivized [15], [16] to share WiFi access. For example, in return Sarah gets bonus points that can be used in future to access internet provided by other crowdsourced services. During a journey on public transport, crowdsourced services can be available at the same transit vehicle, or along the way between two stations. Therefore, the WiFi coverage is regarded as a second layer of service offered on top of the journey planning service. Unlike the work proposed in [11], we do not dispose of any information related to the availability of crowdsourced services at a given location and given time. In addition, to ensure maximum connectivity during the journey, we consider the presence of group of crowdsourced services along a journey plan as a key QoS parameter to select the optimal plan. Indeed, the presence of group of service providers maximizes the chance of failover service availability in case a service becomes unsatisfactory.

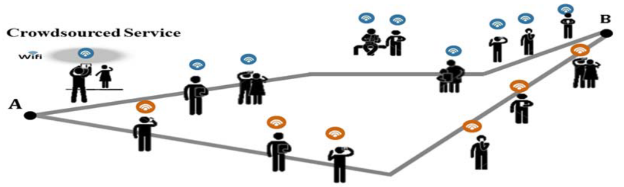

FIGURE 1. Crowdsourced WiFi coverage: Sarah has two candidate optimal journey plan to travel from A to B.

B. JOURNEY PLANNING FRAMEWORK

Fig. 2 illustrates our framework. Given a set of optimal plans, our objective is to determine the best public transport journey plan in terms of WiFi coverage. This set consists of journey plans that result from applying *STA** algorithm [10], a composition approach of line segment services. The set of optimal plans satisfies user requirement such as maximum travel time, maximum waiting time at a station, etc. A line segment is an abstraction of sensors embedded in the transit vehicle on the cloud. It is characterized by its spatiotemporal attributes: source and destination points and departure and arrival time in addition to its QoS such as travel time. By applying *STA**, a composition of line segment services is obtained and results in a set of optimal linear plans that fulfill user requirements such as the preferred travel time. In this framework, the crowdsourced WiFi coverage is modeled in space and time. It has a set of functional and non-functional attributes. We discuss in section III the details of this model. We also dispose of the track records of each crowdsourced service in form of trajectories, i.e a sequence of time stamped geolocations. Using the spatiotemporal model and the trajectory log, we use a probabilistic approach to determine the availability of services in a particular location at a particular time. The first step consists of applying a reference spot detection algorithm to determine the most frequently visited areas. Then, we analyze the periodicity with respect to each reference spot. Given the set of optimal plans, we determine the relevant reference spots crossed by the line segments. To ensure the availability of failover service, we propose a new abstraction of services as a group. The grouping is conducted by deriving the set of intersecting reference spots. A group of crowdsourced services is characterized by its QoS. Finally, we propose an algorithm that takes the set of journey plans, the availability of services and their associated groups and derives the best journey plan in term of WiFi coverage.

In the next sections, we give details of each component of the framework.

III. GROUP OF CROWDSOURCED SERVICES FOR JOURNEY PLAN

In this section, we present the abstraction of WiFi coverage provided by the crowd as a service. We also detail the abstraction of crowdsourced services as a group.

A. CROWDSOURCED WiFi COVERAGE AS A SERVICE

We present in the following, the spatiotemporal model of crowdsourced service S. We adopt the model proposed in [11]. The abstraction of crowdsourced WiFi coverage in space and time is illustrated in Fig. 3. It represents the crowdsourced sensor (typically a smartphone) with its functional and non-functional attributes [11]:

- ID: unique identifier.
- Sensors: Set of sensors. We suppose that the service consists of one sensor at location *loc* and sensing area of radius R_s.
- Space-time: spatio-temporal domain of S. The space is described by a square representing the minimum bounding box of the coverage area as illustrated in Fig. 3. The time is a tuple (t_s, t_e) where t_s and t_e are the start and end time of the service availability at the current sensing area.
- Trajectory: a set of historical K geospatial time stamped locations $(x_i, y_i, t_i)_{1 \leq i \leq K}$ representing the itinerary of S where x_i, y_i and t_i are respectively, the latitude, longitude and timestamp.

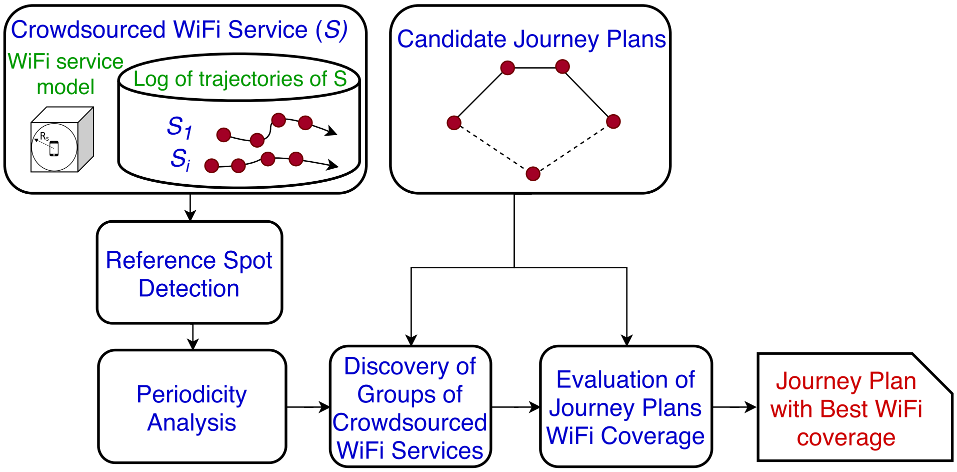
Crowdsourced WiFi Service (S)
WiFi service model
Log of trajectories of S
S_1
S_i
R_s
Candidate Journey Plans
Reference Spot Detection
Periodicity Analysis
Discovery of Groups of Crowdsourced WiFi Services
Evaluation of Journey Plans WiFi Coverage
Journey Plan with Best WiFi coverage

FIGURE 2. Summary of the proposed probabilistic approach for maximizing travel journey WiFi coverage using mobile crowdsourced services.

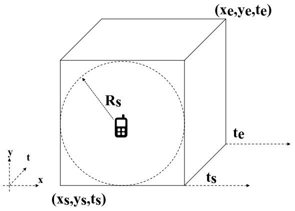

FIGURE 3. Crowdsourced sensor service.

- Set of functionalities F: describes the offered service e.g. providing WiFi access.
- Set of QoS properties Q = $(q_1, q_2, \ldots, q_n)$: where q_i is a QoS property.

A crowdsourced service is characterized by its QoS. We adopt the quality model proposed by Neiat *et al.* [11]. Taking into account the type of the offered service, i.e. WiFi access, the model describes the WiFi signal coverage and the service usage:

Coverage cov: It is associated to the coverage of the WiFi signal with respect to the position from a linear plan. The stronger the signal is, the better the coverage is. Given a confident radius R_c and coverage radius R_s with $R_c < R_s$, the coverage cov is expressed as follows:

$$cov = \begin{cases} 1 & \\ & if \quad d(P, loc) \leq R_c \\ exp\Big(- k(d(P, loc) - R_c)\Big) & \\ & else \end{cases} \tag{1}$$

where $d(P, loc)$ is the perpendicular distance between the linear plan and the location loc of the crowdsourced service. k is a system related constant.

Capacity cap: It reflects the amount of available bandwidth for each WiFi access request:

$$cap = \frac{TR}{NCR} \tag{2}$$

It is the ratio of the total available bandwidth TR of S over the number of users NCR requesting access to the service. In other words, cap describes the crowdsourced service usage.

B. GROUP OF CROWDSOURCED SERVICES

Motivated by the requirement to ensure continuous WiFi coverage during a journey, we propose to take the abstraction of services into a second level and consider the group of crowdsourced services. Indeed, the set Q has fluctuating attributes, i.e. q_i value varies across time. This fluctuation imposes ensuring a failover mechanism to ensure the continuation of the WiFi coverage during the journey. By considering the grouping of services, we can mitigate the risk of disconnection in case of service failure since we can favor joining a group of crowdsourced services with higher availability or failover capacity. This abstraction allows to assign to a group of services a set of QoS. We describe in the following this abstraction and the quality model of a group of crowdsourced WiFi service.

A group is defined with respect to the intersection of particular areas called reference spots which will be further detailed in section IV.

The quality model of a group of crowdsourced services is inherited from the quality model of its members. it is characterized by its confidence, failover capacity, coverage and capacity:

- **Group confidence** $conf$: We define the confidence of a group of crowdsourced services as the ratio of the sum of probabilities of availability of crowdsourced services over the total number of crowdsourced services. We detail in the next section the approach to calculate this particular probability.
- **Group failover capacity** q_{fo}: It is the number of crowdsourced services, members of the group. Each individual service represents a failover in case a service is no longer available or delivers unsatisfactory QoS.

- **Group coverage** q_{cov}: At a given timestamp, multiple crowdsourced services, members of a group, can intersect a linear plan. Therefore, q_{cov} is defined as the normalized sum of coverages of crowdsourced services:

$$q_{cov} = \sum_{i} \frac{\max\left(cov_{S_i}\right) - cov_{S_i}}{\max\left(cov_{S_i}\right) - \min\left(cov_{S_i}\right)} \tag{3}$$

- **Group capacity** q_{cap}: It is defined as the normalized sum of capacities of the crowdsourced services, members of the group:

$$q_{cap} = \sum_{i} \frac{\max\left(cap_{S_i}\right) - cap_{S_i}}{\max\left(cap_{S_i}\right) - \min\left(cap_{S_i}\right)} \tag{4}$$

- **Overall quality** q_{tot}: It is the sum of the weighted aforementioned qualities:

$$q_{tot} = w_1 \cdot conf + w_2 \cdot q_{fo} + w_3 \cdot q_{cov} + w_4 \cdot q_{cap} \tag{5}$$

  where w_i is a weight that reflects the importance of each quality parameter.

IV. PROBABILISTIC PERIODIC BEHAVIOR FOR CROWDSOURCED SERVICE AVAILABILITY

In this section, we describe the mechanism used to determine the availability of services at a given space and time using a reference spot detection approach and thus calculating *conf*, a key QoS parameter to select a group of crowdsourced service.

Realistically, we do not dispose of information related to crowdsourced service availability. Therefore, it is essential to develop approaches that allow finding the available service with degree of certainty.

Periodicity is an intrinsic property of human being. We frequently visit the same places as part of our daily routine. Thus, to discover crowdsourced services along a linear plan, we analyze the historical movement of services to determine the frequently visited areas, called also reference spots. To achieve this goal, we propose a modified version of the approach proposed by Li *et al.* [13], [14] where we use a Jensen-Shannon divergence to reduce computation complexity while maintaining accurate performance. The reference spot detection pipeline is illustrated in Fig. 4.

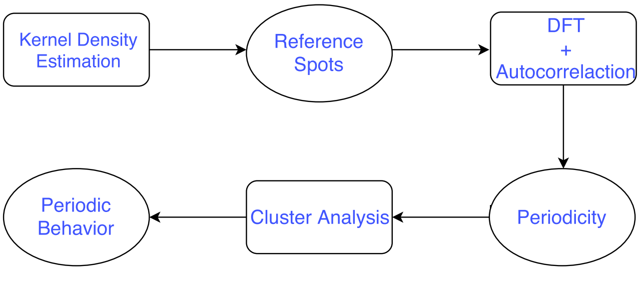

FIGURE 4. The pipeline for mining the periodic movement behaviors of services.

Given the set of optimal journey plans and the set of historical service trajectories, the service availability prediction process uses a kernel density estimation to find the reference spots of crowdsourced service. Then, the periodicity of each crowdsourced service with respect to each reference spot is derived. At runtime, to select the optimal linear plan, we focus on the reference spots crossed by the linear plans, find the associated crowdsourced services and their periodicities and determine the probability of being at the reference spot at a particular time during the period. In the following, we give details of each step.

A. DISCOVERING REFERENCE SPOTS

A reference spot is a dense location frequently visited in the movements. To discover such locations, we partition the area of study into a $w \times h$ grid and compute the density of each cell. Such approach is widely adopted in studying animal movement [17], [19]. Indeed, if an animal exhibits high activity at a place, it represents its home or nest. To estimate the density of each cell c, we use the bivariate normal kernel:

$$f(c) = \frac{1}{n\gamma^2} \sum_{i=1}^{n} \frac{1}{2\pi} exp\left(\frac{d(c, p_i)^2}{2\gamma^2}\right) \tag{6}$$

where $d(c, p_i)$ is the distance between the cell c and the location p_i from the trajectory of the crowdsourced service of length n. γ is the bandwidth of the kernel which is approximated as follows [17]:

$$\gamma = \frac{n^{-1/6}}{2} \sqrt{\sigma_x^2 + \sigma_y^2} \tag{7}$$

σ_x and σ_y are the standard deviations in x and y directions. The density estimation allows finding the reference spots by joining the cell of equal density estimation given some density threshold th. The bigger the threshold is, the larger the size of the reference spot is.

B. FINDING PERIODICITY

To determine the periodicity of a crowdsourced service with respect to a reference spot, the corresponding movement sequence is converted into a binary sequence b_i where $b(i) = 1$ if the crowdsourced service is at the reference spot and 0 otherwise. To calculate the period of the binary sequence, we use a joint Autocorrelation and Discrete Fourier (DFT) Transform approach [18]. The joint analysis is motivated by the following: first, autocorrelation provides accurate estimation of small and large periods. However, it is difficult to set the significance threshold for important periods. On the other hand, with low frequency, DFT provides a poor estimation of large periods as it suffers from the low resolution problem not to mention the false positive generated in the periodogram. Therefore, a joint DFT and autocorrelation analysis can lead to high accuracy estimation of the binary sequence period.

Discrete Fourier Transform: The normalized DFT of sequence $b(i)$, $i = 1 : N$ is a sequence of complex numbers $B(f)$:

$$B(f_{k/N}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} b(i) exp\left(\frac{-j2\pi ki}{N}\right) \quad k = 1 \ldots N \tag{8}$$

k/N is the frequency captured by each coefficient. The period is defined as the inverse of the frequency. Therefore, by identifying the frequencies that hold most of the energy, we can determine the most dominant periods. Given the coefficients of the sequence, the periodogram P is defined as the square of each coefficient: $P(f_{k/N}) = ||B(f_{k/N})||^2$. To identify the most dominant frequency, a thresholding approach is adopted. Once the dominant frequencies are identified, mapping them to time domain is required. A single coefficient corresponds to the period range $[\frac{N}{k}, \frac{N}{k-1})$ in time domain. In order to accurately determine the period, we use circular autocorrelation.

Circular autocorrelation Circular autocorrelation $R(\tau)$ examines the similarity of a sequence to its previous values for different lags τ:

$$R(\tau) = \frac{1}{N}\sum_{i=1}^{N} b(\tau)b(i+\tau) \tag{9}$$

Given a period range $[t_1, t_2)$ obtained using the periodogram analysis, we check if there is a peak in $\{R(t_1), R(t_1+1), \ldots, R(t_2-1)\}$ by quadratic function fitting. If the obtained fitted function is concave, it reflects the presence of a period $t^* = argmax_{t_1 \leq t < t2} R(t)$.

C. PROBABILISTIC PERIODIC BEHAVIOR

The availability of a crowdsourced service at a given reference spot and given time is modeled using a probabilistic model. For example, crowdsourced service #1 is located at reference spot #1 at 10 AM with a probability 0.8. To determine the probability, let $RF = rf_1, rf_2, \ldots, rf_r$ be the set of reference spots of common period T. Given the sequence of historical locations $loc_1, loc_2 \ldots$, the sequence of presence at the reference spots is $q_1, q_2, \ldots, q_n$ where $q_i = j$ if the crowdsourced service location loc_i is at reference spot j. The sequence q_i is further divided into $m = \frac{n}{T}$ segments set $\mathcal{I} = \{I_i\}$. Let the timestamps set be $\mathcal{T} = \{t_1 \ldots t_T\}$. We denote by $\mathcal{M}$ the probability matrix of size $r \times T$ where each element represents the probability of a segment I_i being at reference spot rf_j at time t_k. Assuming an independent categorical distribution prior, the probability that maximizes the likelihood represents the best generative model that reflects the probability of crowdsourced service being at reference spot j at time t_k. Each element of $\mathcal{M}$ is expressed as:

$$p(I_i \ in \ rf_j \ at \ time \ t_k) = \frac{\sum_{I_i} \mathbb{1}_{I_i \ in \ rf_j \ at \ time \ k}}{|\mathcal{I}|} \tag{10}$$

where $\mathbb{1}$ is the indicator function, that is $p(I_i \ in \ rf_j \ at \ time \ t_k)$ is the relative frequency of reference spot rf_j at time t_k over all segments in $\mathcal{I}$. A periodic behavior B [13] is thus defined as a pair $(T, \mathcal{M})$ where T is the period and $\mathcal{M}$ is the probability with respect to the set of reference spots associated to the crowdsourced service.

Now, given a set of segments, we need to find the set of segments generated by the same periodic behavior. In [13], authors suggested using the agglomerative hierarchical clustering approach [20] to group these segments where each group represents a periodic behavior. The choice of distance in in this particular type of clustering greatly influences the final data partition. Authors proposed to use Kullback-Leiber (KL) divergence [21], [22] as a distance between two distributions. However, such divergence measure is reported to be not a good estimator even in presence of high instances drawn from the distributions [23]. In addition, KL divergence is not symmetric thus not practical in many applications where computation complexity is of paramount importance. Furthermore, KL is not defined when the probability is equal to 0. To cope with these issues, Li *et al.* [13] proposed to use background variable sampled from uniform distribution and smoothed with a positive parameter to solve this problem. For these particular reasons, we propose to use Jensen-Shannon (JS) divergence. JS is used to overcome the aforementioned drawbacks. Indeed, JS is symmetric, finite and semi-bounded. Consequently, JS allows further reduction in computation complexity and does not require any smoothed background variable. Given two distributions P_1 and P_2, JS is defined as:

$$JS(P_1, P_2) = \frac{1}{2}KL(P_1, \hat{P}) + \frac{1}{2}KL(P_2, \hat{P}) \tag{11}$$

where $\hat{P} = \frac{P_1+P_2}{2}$ and $KL(P_1, P_2)$ is defined as:

$$KL(P_1, P_2) = -\sum_x P_1(x) log\Big(\frac{P_1(x)}{P_2(x)}\Big) \tag{12}$$

For better understanding, let us consider the following statistical analysis. From [13], given a set of segments $\mathcal{I}$ generated by a distribution P_1, we have:

$$KL(P_1, P_2) = -H(P_1) - \frac{1}{|\mathcal{I}|} log(\mathcal{P}(\mathcal{I}, P_2)) \tag{13}$$

where $H(P_1)$ is the entropy of P_1 and $\mathcal{P}(\mathcal{I}, P_2)$ is the probability that the whole set of segments is generated by a distribution P_2. Using Eq. 11, we have:

$$\begin{aligned} JS(P1, P2) &= \frac{1}{2}KL(P_1, \hat{P}) + \frac{1}{2}KL(P_2, \hat{P}) \\ &= \frac{1}{2}\Big(- H(P_1) - \frac{1}{|\mathcal{I}|}log(\mathcal{P}(\mathcal{I}, \hat{P}))\Big) \\ &\quad + \frac{1}{2}\Big(- H(P_2) - \frac{1}{|\mathcal{I}|}log(\mathcal{P}(\mathcal{I}, \hat{P}))\Big) \\ &\quad \times \frac{1}{2}\Big(- H(P_1) - H(P_2)\Big) - \frac{1}{|\mathcal{I}|}log(\mathcal{P}(\mathcal{I}, \hat{P})) \end{aligned} \tag{14}$$

Assuming $H(P_1)$ and $H(P_2)$ are constant, JS assesses how likely the set of segments are generated by a mixture of distributions. This provides better modeling of human behavior as a periodic behavior can be influenced by another one. For example, a periodic behavior consists of being at office till 5 PM during weekdays and going for lunch at nearby restaurant during the break time from 12 PM to 1 PM. In terms of computation complexity, the agglomerative hierarchical clustering involves calculating pairwise distance matrix. Given the non-symmetric property of KL, this requires $m(m-1)$ operations while it only requires $\frac{m(m-1)}{2}$ operations using JS given its symmetric property. Once the set of segments are grouped, each cluster represents the set of segments with the same periodicity. Finally, we can calculate the confidence quality $conf$ as:

$$conf = \frac{\sum_{S_i} p(I_i \ in \ rf_j \ at \ time \ t_k)}{|S_i|} \tag{15}$$

V. OPTIMAL JOURNEY PLAN SELECTION

We present in the following the proposed algorithm to select the best linear plan from source to destination. Algorithm 1 describes the details of the algorithm. The selection process is as follows: for each line segment, we find the set of intersecting reference spots during the journey from the source point to destination point (Line 5). For each intersecting reference spot, we identify its associated crowdourced services, i.e. the services that frequently visit this particular spot. A reference spot is a collection of reference points forming an area for which we can calculate the convex hull i.e. the small set that contains the reference points.. To derive the group of crowdsourced service, we identify the intersecting convex hulls of the other reference spots. Thus, the group of crowdsourced services are characterized by their intersecting reference spots. Next, for each crowdsourced service, member of the group, we calculate its capacity and coverage (Line 7-8). This allows us to calculate for each group its associated qualities (Line 11-14). Finally, the overall WiFi coverage quality of the line segment with respect to a reference spot is calculated as follows:

$$q_{WiFi} = \sum_i q_{tot}/L \tag{16}$$

where L is the number of line segments composed to form the linear plan.

VI. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of our framework. To the best of our knowledge, there are no available spatiotemporal services data to assess our approach. Therefore, we use publicly available mobility data. We assume that under certain moving patterns such as walking, the obtained traces would correspond to crowdsourced hotspot services. In the following, we present the data, how it has been processed and the evaluation performance of the framework.

A. DATA

We conduct our experiments on the Geolife GPS trajectory dataset [24]. Traces were collected by Microsoft Research Asia under the Geolife project. 178 users were tracked for four years from April 2007 to October 2011 over 30 Chinese cities, USA and Europe, although the majority of the traces was collected in Beijing, China. Users include University students, Microsoft employees, government staff and employees of other companies. Each track is a set of time stamped geolocation (latitude and longitude). A total of 17621 trajectories were collected, 91% of which have a sampling rate of 1 to 5 seconds or 5 to 10 meters. Users were tracked while conducting different activities: walk, bike, bus, car&taxi, train and plane. More than 42% (5,436 hours of 12,856 hours) of the labeled traces correspond to the walking activity. We assume that the walking traces correspond to crowdsourced WiFi coverage service. To avoid the sparsity of walking records, we also assume that data are sampled at consecutive times without any missed samples as it is challenging to detect period of sparse data [25] .

Algorithm 1 Journey Plan Selection Based on Crowdsourced WiFi Coverage

```
1:  Input  Set of optimal linear plans, Set of reference spots $RF$
2:  Output: Best linear plan
3:  for each linear plan do
4:      for each line segment $ls_i$ do
5:          $RF_{int} = RF \cap ls_i$
6:          for each $rf_i \in RF_{int}$ do
7:              Calculate $cov(S_{rf_i})$ (Eq. 1).
8:              Calculate $cap(S_{rf_i})$ (Eq. 2).
9:              $\overline{RF}_{int} = rf_i \cap RF$
10:             $Group = \{\}$
11:             for each $rf_j \in \overline{RF}_{int}$ do
12:                 Calculate $cov(S_{rf_j})$ (Eq. 1).
13:                 Calculate $cap(S_{rf_j})$ (Eq. 2).
14:                 Insert $S_{rf_j}$ in $Group$
15:             Calculate $q\_cov(Group)$ (Eq. 3)
16:             Calculate $q\_cap(Group)$ (Eq. 4)
17:             Calculate $conf$ (15)
18:             Calculate $q\_fo$
19:             Calculate $q\_tot$ (Eq. 5)
20:     Calculate $q\_WiFi$ (Eq. 16)
21: Best linear plan = argmax($q\_WiFi$)
```

To efficiently extract the reference spots, we convert the latitude and longitude traces to Cartesian axis with respect to a reference point. Without loss of generality, we choose Microsoft China Research And Development Group Headquarters Building 2 in Beijing as the reference point. It is the point whose latitude $lat_r = 39.980888$ and longitude $long_r = 116.310160$. Assuming that the earth radius $R = 6371 \ 10^3 \ m$, the new longitude $long_n$ is calculated using the observed longitude $long_o$ and latitude lat_o:

$$long_n = \frac{R \times \pi}{180} \times \left(long_o - long_r\right) \times cos\left(lat_o \times \frac{\pi}{180}\right) \tag{17}$$

The new latitude lat_n is derived from the observed latitude lat_o:

$$lat_n = \left(lat_o - lat_r\right) \times \frac{R \times \pi}{180} \tag{18}$$

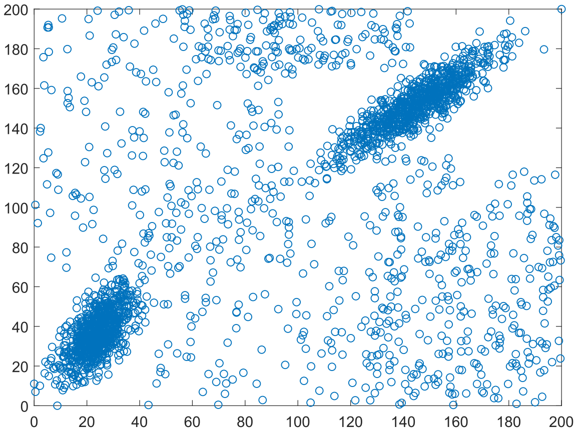

FIGURE 5. Simulated movement with two potential reference spots.

For the set of linear plans, we establish a scenario in which we receive a request to travel from China Academy of Sciences Institute of Automation (latitude = 39.980197, longitude = 116.333305) to Peking University (latitude = 39.986914, longitude = 116.305880). We assume user preferences are relaxed in terms of journey plan (long waiting time and walking distance are accepted) to include maximum candidate linear plans. By applying *STA** algorithm, four linear plans are available: three direct linear plans namely Bus 641, 333 and 913, and one linear plan requires a connection between two bus services: 466 and 608. For Bus 641 plan, it includes walking 4 min to the departure station and 6 min to the destination from the arrival station. Bus 333 service includes 10 min walking to the departure station and later 1 min to the destination. The third plan requires 11 min walking to the departure station and later 7 min to the destination. For the fourth plan, it requires 5 min walking to the departure station of Bus 466, then 1 min walking to transfer to Bus 608, waiting 9 min for Bus 608 arrival and finally 1 min walking to the destination. To evaluate the accuracy of the reference spot detection based on *JS* divergence, we use a ground truth synthetic data depicted in Fig.5. These data simulate movement on hourly basis in 200 × 200 area of study and with well-established behavior: two reference spots with same period: T = 24. We deliberately set the traces as follows: between 0:00 AM and 8 AM, the movement occurs at Reference Spot 2 while it occurs at Reference Spot 1 between 5:00 PM and 0:00 AM. In between, the movement randomly occurs in the area of study.

B. REFERENCE SPOTS DETECTION

First, we assess the performance of the reference spot detection based on *JS* against the ground truth dataset. Next, we analyze the periodic behavior of users using the Geolife dataset traces.

1) PERFORMANCE EVALUATION

Fig. 6 depicts the reference spots associated to the synthetic ground truth data. The *JS*-based reference spot algorithm is able to detect both two reference spots. We illustrate in Fig. 7 the density estimate of the movement. The circular autocorrelation shown in Fig. 8 reveals a pick that corresponds to a period $T = 24$. The *JS* based periodic behavior is depicted in Fig. 9. For the first 8 hours, the movement is essentially located at Reference spot 1 with high probability: $p(I_i \ in \ rf_2 \ at \ time \ t_k) > 0.95$ $(1 \leq t_k \leq 8, i \in [1, 8])$. We notice no activity that occurs at Reference spot 1. During the last 8 hours of the day, the activity occurs in Reference spot 2 with high probability (> 0.98). We notice no activity during this time at Reference spot 2. During the rest of the day, the activity occurs in unknown places. This behavior perfectly coincides with the simulated behavior deliberately established in the synthetic data.

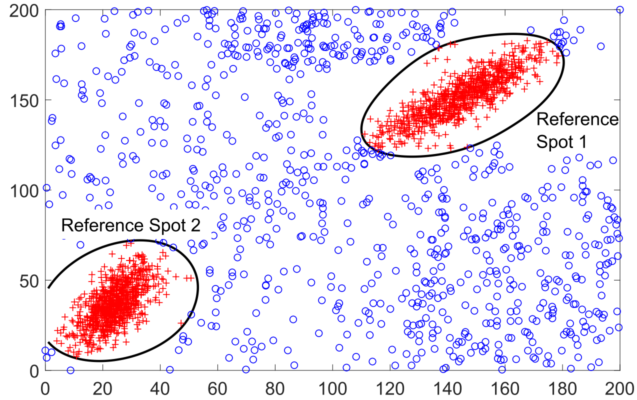

FIGURE 6. Two reference spots are detected.

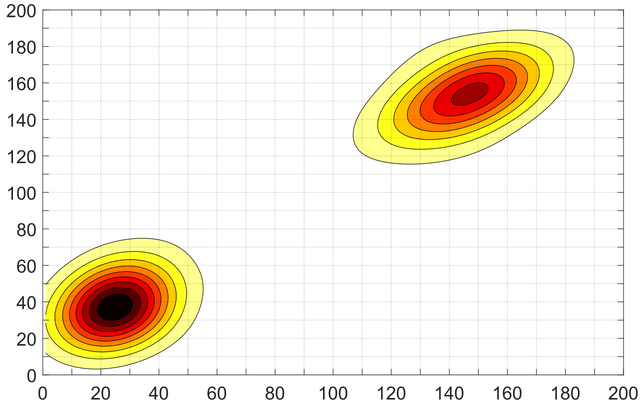

FIGURE 7. Contours of the density estimate.

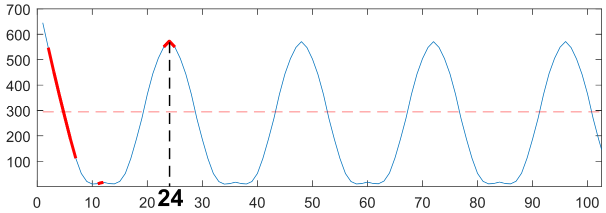

FIGURE 8. Circular autocorrelation of the synthetic movement.

We further evaluate the time complexity of the reference spot detection based on both *KL* and *JS* approaches. Fig. 10 depicts the execution time variation with respect to the

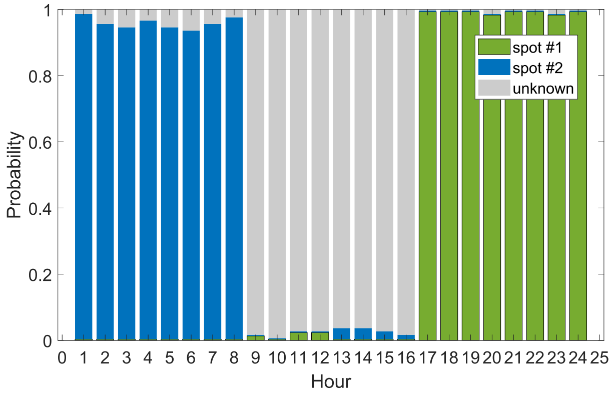

FIGURE 9. Periodic behavior during the period $T = 24$.

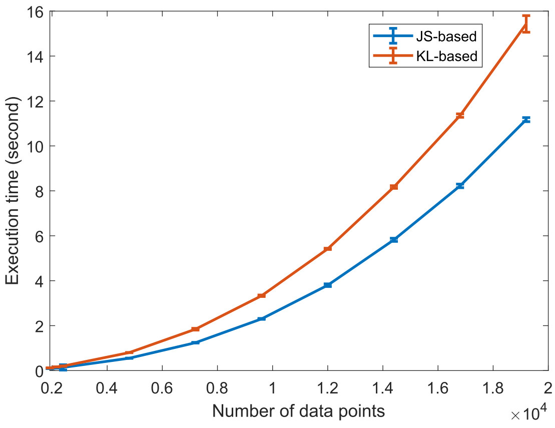

FIGURE 10. Execution time with respect to the number of points.

number of points in the synthetic data. The *JS* based approach requires less time compared to the *KL* based one thanks to the symmetry of the divergence measure.

2) GEOLIFE DATA ANALYSIS

We randomly choose two traces for analysis and identification of reference spots and the potential periods of visiting. We illustrate in Fig. 11 the walk traces of user #105. A visual inspection reveals that we have four potential reference spots since there are four dense areas of traces. Fig. 12 illustrates the reference spots detected. In fact, based on his pattern, user #105 frequently visits spots 1 and 3. An investigation of the map shows that reference spot 1 corresponds to an area within Renmin University of China in Haidian District of Beijing. Thus, it is very likely that these traces correspond to a university student. The periodicity of this spot is equal to 49 hours, i.e. user #105 participates in this experiment while walking every 2 days. Reference spot 3 corresponds to a business district of shopping malls and attractions around CBD Historical and Cultural Park. User #105 visits this area every 162 h i.e. on a weekly basis. Fig. 13 shows the probability of availability with respect to reference spot 1. We notice that in time interval [1, 10], user is located in reference spot 1 with a probability ≈ 0.45. This probability gets lower and lower till $t = 40$. The 'Unknown' probability reflects

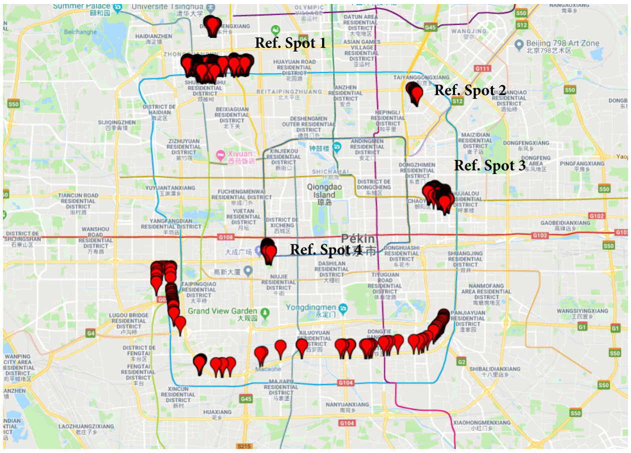

FIGURE 11. Walking traces of user #105.

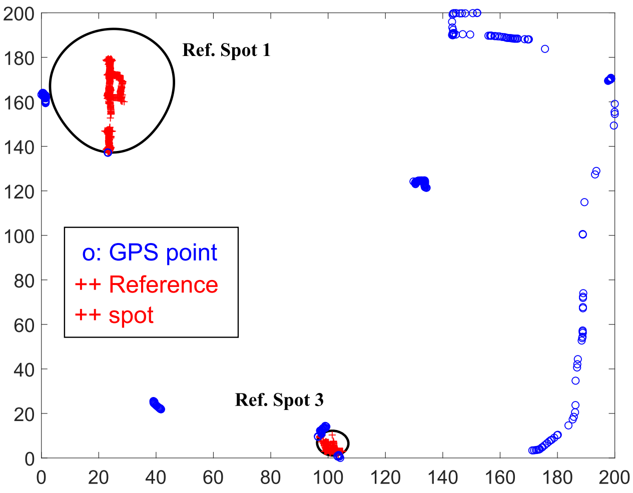

FIGURE 12. Reference spots of user #105. Points are normalized.

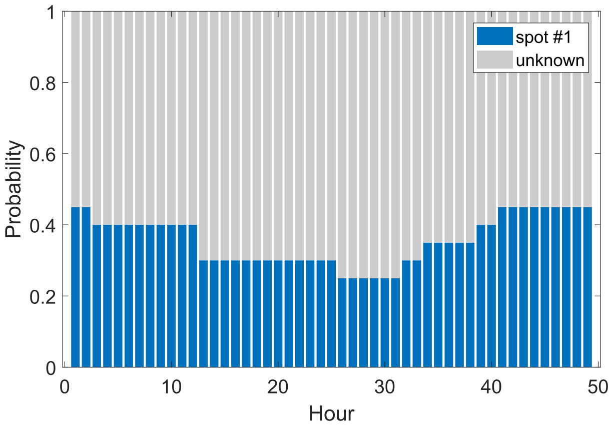

FIGURE 13. Behavior of user #105 during the period of T = 49.

the availability in other reference spots (reference spot 3 for instance.). The behavior with respect to reference spot 3 is depicted in Fig. 14. We notice that most of the activities at reference spot 3 occur at the end of the period.

We further provide analysis for user #125. Fig. 15 illustrates the GPS traces. The algorithm reveals two reference spots as illustrated in Fig. 16. By investigating these spots

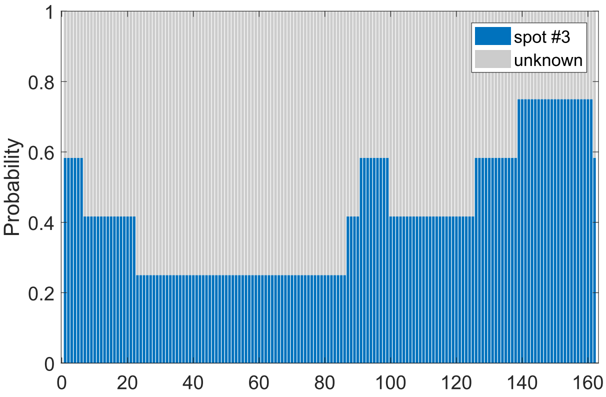

FIGURE 14. Behavior of user #105 during the period of T = 162.

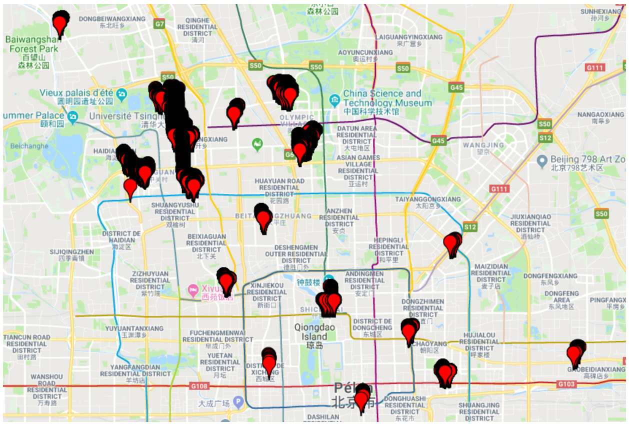
DONGBEIWANGXIANG
Baiwangshan Forest Park
Vieux palais d'été
Université Tsinghua
China Science and Technology Museum
OLYMPIC
DATUN AREA RESIDENTIAL DISTRICT
ASIAN GAMES VILLAGE RESIDENTIAL DISTRICT
WANGJING
Beijing 798 Art Zone
HUAYUAN ROAD RESIDENTIAL DISTRICT
SHUANGYUSHU RESIDENTIAL DISTRICT
TAIYANGGONGXIANG
HEPINGLI RESIDENTIAL DISTRICT
MAIZIDIAN RESIDENTIAL DISTRICT
DONGFENGXIANG
Qiongdao Island
DONGZHIMEN RESIDENTIAL DISTRICT
CHAOYANG
GAOBEIDIANXIANG
Pékin

FIGURE 15. Walking traces of user #125.

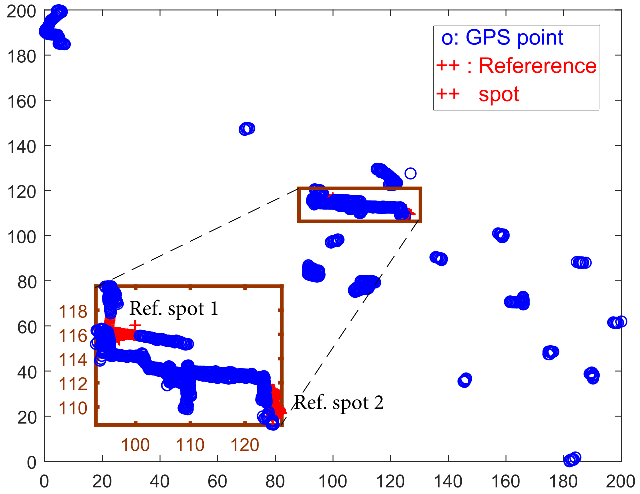

FIGURE 16. Reference spots of user #125. Points are normalized.

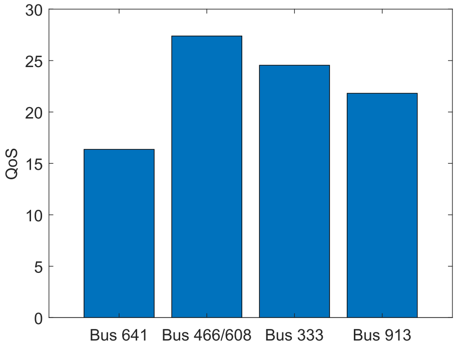

FIGURE 17. Quality of WiFi coverage of the candidate linear plans.

on a map, we discover that spot 1 corresponds to Tsinghua University located in Haidian District, Beijing China, specifically Tsinghua Park. Thus, it is very likely that these traces correspond to a university student. User #125 frequently visits and walks across this reference spot every 296 hours or 12 days and 8 hours. Reference spot 2 corresponds to a residential area. User #125 visits and share his information in this area with a period of 91 hours or 3 days and 19 hours.

C. OPTIMAL JOURNEY PLANNING

In the following, we conduct the scenario under which, we have a request to travel from China Academy of Sciences Institute of Automation to Peking University as detailed in section VI-A. Four potential linear plans are found. Ignoring the desire for WiFi access, one would automatically pick the first plan as it requires less travel time. For simulation purposes, we set $R_c = 3m$, $R_s = 10m$ and the constant $k = 0.5$. We sample the capacity of each crowdsourced service from uniform distribution. To determine the intersected reference spots by the linear plan, we define for each reference spot its boundary obtained by its convex hull. The convex hull is the smallest convex set containing all points in the reference spots. Thus, finding the intersecting lines become a simple geometrical problem. We assume that the crowdsourced service is located at the center of its reference spot which enables the calculation of the coverage quality cov. We also set equal contribution of qualities of a group i.e. $w_i = 1$ for $i = 1 \ldots 5$ (Eq. 5).

Our algorithm findings are illustrated in Fig. 17. Results show that although the journey plan involving both Bus 466 and Bus 608 requires a transit between two vehicles, it exhibits the best quality in terms of WiFi coverage. Bus 641 service, although has the less travel time, exhibits the least WiFi connectivity. We further evaluate the time complexity of the proposed algorithm. Fig. 18 shows that for all available linear plans, the algorithm requires 10 ms to calculate the WiFi quality of each linear plan which includes finding the intersecting reference spots, the groups of crowdsourced services and the corresponding quality.

D. DISCUSSION

The proposed framework builds a second service layer on top of the journey plan service to provide connectivity during the journey. The framework can be characterized as best effort as it does not ensure connectivity all time. Indeed, the linear plan should intersect at least one reference spot associated to a crowdsourced service. However, with higher participation

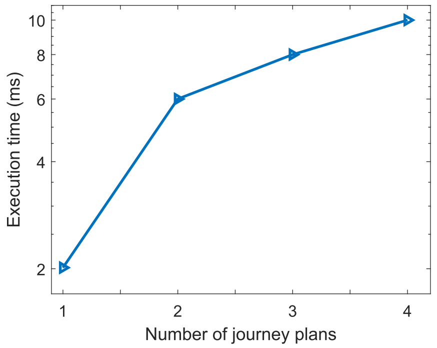

FIGURE 18. Execution time versus number of linear plans.

from the crowd, the area of study can be fully covered by the crowdsourced services. In addition, one may include static WiFi services such as the ones provided in public areas, coffee shops, etc. From periodic behavior perspective, the *JS*-based approach is able to accurately mine the periodicity of the movement as confirmed by the synthetic data experiment with less time complexity compared to the original approach. One downside of this approach is the requirement of mining continuously sampled time stamped locations. In addition, with presence of sparse data, mining periodicity becomes a challenging task. As shown by the Geolife experiments, the probability of availability at a given reference spot in some cases is neither high nor low which introduces uncertainty about the availability of crowdsourced services. This is explained by the complexity of human behavior as it is much more complex to model in comparison to animals for example whose behavior is easier to capture.

VII. CONCLUSION

This paper proposed a probabilistic framework to select the optimal journey plan based on the quality of WiFi offered by crowdsourced services along the journey. The probabilistic model uses the periodic movement behaviors of services to derive the availability of crowdsourced services. This model is based on Jensen-Shannon divergence rather than Kullback-Leiber to reduce the computational cost as Jensen-Shannon is symmetric. Experimental results on synthetic data demonstrated the accuracy of the Jensen-Shannon-based approach in mining the periodic behavior. We also established a scenario using real-world GPS traces and public transport scenario. In future work, we will augment the proposed framework to take into account static WiFi services such as the ones available in public attractions areas and coffee shops. We will also study the case of sparse GPS trace data for mining the periodic movement behaviors.

ACKNOWLEDGMENT

The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] A. Ben Said, A. Erradi, A. G. Neiat, and A. Bouguettaya, “Mobile crowdsourced sensors selection for journey services,” in *Proc. Int. Conf. Service Oriented Comput.*, 2018, pp. 463–477.
[2] X. Chen, E. Santos-Neto, and M. Ripeanu, “Crowdsourcing for on-street smart parking,” in *Proc. 2nd ACM Int. Symp. Design Anal. Intell. Veh. Netw. Appl.*, 2012, pp. 1–8.
[3] Y. F. Dong, S. Kanhere, C. T. Chou, and N. Bulusu, “Automatic collection of fuel prices from a network of mobile cameras,” in *Proc. Int. Conf. Distrib. Comput. Sensor Syst.*, 2008, pp. 140–156.
[4] T. Kandappu, A. Misra, D. Koh, R. D. Tandriansyah, and N. Jaiman, “A feasibility study on crowdsourcing to monitor municipal resources in smart cities,” in *Proc. Web Conf.*, 2018, pp. 919–925.
[5] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, “Urban computing with taxicabs,” in *Proc. 13th ACM Int. Conf. Ubiquitous Comput.*, 2011, pp. 89–98.
[6] M.-P. Pelletier, M. Trépanier, and C. Morency, “Smart card data use in public transit: A literature review,” *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 4, pp. 557–568, 2011.
[7] Y. Lu, A. Misra, W. Sun, and H. Wu, “Smartphone sensing meets transport data: A collaborative framework for transportation service analytics,” *IEEE Trans. Mobile Comput.*, vol. 17, no. 4, pp. 945–960, Aug. 2017.
[8] A. G. Neiat and A. Bouguettaya, *Crowdsourcing of Sensor Cloud Services*. Springer, 2018.
[9] A. Ben Said, A. Erradi, A. G. Neiat, and A. Bouguettaya, “A deep Learning spatiotemporal prediction framework for mobile crowdsourced services,” *Mobile Netw. Appl.*, vol. 24, no. 3, pp. 1120–1133, 2018.
[10] A. G. Neiat, A. Bouguettaya, T. Sellis, and Z. Ye, “Spatio-temporal composition of sensor cloud services,” in *Proc. Int. Conf. Web Service*, Jun. 2014, pp. 241–248.
[11] A. G. Neiat, A. Bouguettaya, and T. Sellis, “Spatio-temporal composition of crowdsourced services,” in *Proc. Int. Conf. Service Oriented Comput.*, 2015, pp. 373–382.
[12] A. G. Neiat, A. Bouguettaya, T. Sellis, and H. Dong, “Failure-proof spatio-temporal composition of sensor cloud services,” in *Proc. Int. Conf. Service-Oriented Comput.*, 2014, pp. 368–377.
[13] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye, “Mining periodic behaviors for moving objects,” in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 1099–1108.
[14] Z. Li, B. Ding, J. Han, and R. Kays, “Mining periodic behaviors of object movements for animal and biological sustainability studies,” *Data Mining Knowl. Discovery*, vol. 24, no. 2, pp. 355–386, 2012.
[15] K. Han, H. Huang, and J. Luo, “Quality-aware pricing for mobile crowdsensing,” *IEEE/ACM Trans. Netw.*, vol. 26, no. 4, pp. 1728–1741, Aug. 2018.
[16] Y. Wang, X. Jia, Q. Jin, and J. Ma, “QuaCentive: A quality-aware incentive mechanism in mobile crowdsourced sensing (MCS),” *J. Supercomput.*, vol. 72, no. 8, pp. 2924–2941, 2016.
[17] B. J. Worton, “Kernel methods for estimating the utilization distribution in home-range studies,” *Ecology*, vol. 70, no. 1, pp. 164–168, 1989.
[18] M. Vlachos, P. S. Yu, and V. Castelli, “On periodicity detection and structural periodic similarity,” in *Proc. SIAM Int. Conf. Data Mining*, 2005, pp. 449–460.
[19] B. J. Worton, “A review of models of home range for animal movement,” *Ecol. Model.*, vol. 38, nos. 3–4, pp. 277–298, 1987.
[20] A. Ben Said, R. Hadjidj, and S. Foufou, “Cluster validity index based on Jeffrey divergence,” *Pattern Anal. Appl.*, vol. 20, no. 1, pp. 21–31, 2017.
[21] S. Kullback, *Information Theory and Statistics*. North Chelmsford, MA, USA: Courier Corporation, 2012.
[22] D. Commenges, “Information theory and statistics: An overview,” 2015, *arXiv:1511.00860*. [Online]. Available: https://arxiv.org/abs/1511.00860
[23] M. Budka, B. Gabrys, and K. Musial, “On accuracy of PDF divergence estimators and their applicability to representative data sampling,” *Entropy*, vol. 13, pp. 1229–1266, Jul. 2011.
[24] Y. Zheng, L. Wang, R. Zhang, X. Xie, and W.-Y. Ma, “GeoLife: Managing and understanding your past life over maps,” in *Proc. Int. Conf. Mobile Data Manage.*, 2008, pp. 211–212.
[25] I. Junier, J. Hérisson, and F. Képès, “Periodic pattern detection in sparse Boolean sequences,” *Algorithms Mol. Biol.*, vol. 5, no. 1, p. 31, 2010.