Spotting Trip Purposes from Taxi Trajectories: A General Probabilistic Model

PENGFEI WANG, University of Chinese Academy of Sciences & Computer Network and Information Center, Chinese Academy of Sciences
GUANNAN LIU, Beihang University
YANJIE FU, Missouri University of Science and Technology
YUANCHUN ZHOU and JIANHUI LI, Computer Network and Information Center, Chinese Academy of Sciences

What is the purpose of a trip? What are the unique human mobility patterns and spatial contexts in or near the pickup points and delivery points of trajectories for a specific trip purpose? Many prior studies have modeled human mobility patterns in urban regions; however, these analytics mainly focus on interpreting the semantic meanings of geographic topics at an aggregate level. Given the lack of information about human activities at pick-up and dropoff points, it is challenging to convert the prior studies into effective tools for inferring trip purposes. To address this challenge, in this article, we study large-scale taxi trajectories from an unsupervised perspective in light of the following observations. First, the POI configurations of origin and destination regions closely relate to the urban functionality of these regions and further indicate various human activities. Second, with respect to the functionality of neighborhood environments, trip purposes can be discerned from the transitions between regions with different functionality at particular time periods.

Along these lines, we develop a general probabilistic framework for spotting trip purposes from massive taxi GPS trajectories. Specifically, we first augment the origin and destination regions of trajectories by attaching neighborhood POIs. Then, we introduce a latent factor, *POI Topic*, to represent the mixed functionality of the regions, such that each origin or destination point in the city can be modeled as a mixture over POI Topics. In addition, considering the transitions from origins to destinations at specific time periods, the trip time is generated collaboratively from the pairwise POI Topics at both ends of the O-D pairs, constituting *POI Links*, and hence the trip purpose can be explained semantically by the POI Links. Finally, we present extensive experiments with the real-world data of New York City to demonstrate the effectiveness of our proposed method for spotting trip purposes, and moreover, the model is validated to perform well in predicting the destinations and trip time among all the baseline methods.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications-Spatial databases and GIS

This work was supported by the National Key Research Program of China under Grants No. 2016YFB1000600 and No. 2016YFB0501900 and the Natural Science Foundation of China under Grant No. 61402435. Dr. Guannan Liu is supported by the Natural Science Foundation of China under Grant No. 71701007 and 71531001. This research was partially supported by University of Missouri Research Board (proposal number: 4991).
Authors' addresses: P. Wang, University of Chinese Academy of Sciences & Computer Network and Information Center, Chinese Academy of Sciences, Beijing, China; email: wpf@cnic.cn; G. Liu, Beihang University, Beijing 100191, China; email: guannliu@gmail.com; Y. Fu (corresponding author), Missouri University of Science and Technology; email: yanjiefoo@gmail.com; Y. Zhou and J. Li, Computer Network and Information Center, Chinese Academy of Sciences, Beijing, China; emails: {zyc, lijh}@cnic.cn.

ACM 2157-6904/2017/12-ART29 $15.00
https://doi.org/10.1145/3078849

General Terms: Design, Implementation Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Human mobility, taxi trajectories, trip purposes, probabilistic model

ACM Reference format:
Pengfei Wang, Guannan Liu, Yanjie Fu, Yuanchun Zhou, Haoyi Xiong, and Jianhui Li. 2017. Spotting Trip Purposes from Taxi Trajectories: A General Probabilistic Model. *ACM Trans. Intell. Syst. Technol.* 9, 3, Article 29 (December 2017), 26 pages.
https://doi.org/10.1145/3078849

1 INTRODUCTION

The rapid development of Internet, mobile, and sensing technologies has accumulated large-scale crowd-sourced human mobility data at a high spatial-temporal resolution from location-aware devices. For instance, with embedded GPS in taxi cabs, people's moving trajectories of taking taxis are recorded. Indeed, the emergence of human mobility data provides an invaluable source for understanding mobility behavioral patterns and further helps urban planning. For instance, some researchers have leveraged public bus smart-card data for planning new stations that could better meet bus service demand; others have utilized human mobility (such as taxi GPS data) for predicting air pollution in big cities.

However, even though the locations of individuals can be tracked from real-time GPS sensing, the trip purposes of these GPS traces are usually unknown, or in other words, we have no idea about people's activities once they leave the taxi. Therefore, there is indeed a dilemma in which we have massive human footprints but lack their activity information, while such activity information can help reveal the purpose of the taxi trips. As a matter of fact, inferring trip purposes can significantly enrich and augment the semantic meanings of trajectories, which can be useful for many applications, for example, improving city planning and governance, supporting transportation engineering, enhancing public safety, understanding social interaction, and ultimately, providing decision-making support for smart-city planning.

Existing methods such as passive methods (e.g., surveys, questionnaires, self-reports) and analytic methods (e.g., Bayesian probability estimation, trajectory clustering) can partially solve the problem of inferring trip purposes. Nevertheless, these may encounter different problems in explaining trip purposes. Passive methods are usually impractical and highly depends on the Willingness-To-Share of mobile users; while traditional Bayesian probability estimation and trajectory clustering are hard to simultaneously strive a balance between accuracy and interpretability. Thus, some recent studies seek to annotate the trips with mobility sequence or other context information. Wu and Li (2016) wisely formulated the problem of annotating the trajectories as a task of predicting the actual venues that the user tend to visit, and the analysis is at personalized level, because the data contains the visiting history of each user. Yan et al. (2013) proposed to model a complete sequence of a moving object, with detail records of the trajectories including the geographical information and the time, and they developed a platform to annotate the trajectories with various background information. These studies rely on the availability of complete trajectory data or background information such as users' profile, check-in history, and so on, however, in most cases, we only have the information about the origins, destinations, and the time of the trip.

Therefore, it highly necessitates a novel analytic framework for spotting trip purposes from large-scale vehicle trajectories without any context information or labels, and moreover, predicting the potential destinations and trip time with the modeling framework. To that end, in this article, the taxi trajectories are modeled in the following perspectives. First, each taxi trajectory consists of a pick-up point and a dropoff point, which can be regarded as a directed Origin-Destination (O-D) pair. Second, considering an O-D pair only indicates the direction of a trip, we augment

the O-D pair with the neighborhood environment of both the origins and destinations, since the neighborhood environment, along with the directions between different functional regions, can reveal the possible human activity types. For instance, the trip purpose from an office region to a residential region is likely to be returning home after work, while the trip purpose from a residential region to an office region is likely to be commuting to work. Third, the neighborhood environment of a specific point can be depicted by the nearby Point of Interests (POI), and the distribution of POIs reflect the mixed functionality of a region. For example, an area composed of many office, shopping, and entertainment buildings is likely to be inferred as Central Business District. Finally, each O-D pair is companied by a particular trip time, which can enhance the inference on the purpose of the trip, because human outdoor activities have temporal patterns. For example, a trip from residential region to restaurant-related region around 7 p.m. is more likely to be dinning purpose compared with the same trips happened in the morning. In summary, trip purposes can be reflected by the inter-correlations of urban functionality between the O-D pairs, which exactly corresponds to the POI distributions in the neighborhood environment; meanwhile, the trip time allows to explore the temporal correlations between the time period and trip purposes.

It is naturally appealing to exploit the urban functionality, directions, and temporal information into a general and unified model for spotting trip purposes, hence we propose a generalized probabilistic model, named POI Link Model, for spotting trip purposes by collectively leverage the POIs of O-D pairs, trajectory directions, and temporal information of human mobility data. To start with, we augment the O-D pairs with the neighborhood environment, that is, the POIs that falls in a predefined neighborhood of the point. In the spirit of traditional topic models, we treat each point as a document and the nearby POIs (e.g., a circle with 200 meters as radius) are words, and we further introduce a latent factor, *POI Topic*, to represent the mixture of POIs that corresponds to certain activities. Then, for each augmented O-D pair, we have POI Topic distributions for both the origins and destinations, respectively, and thus the augmented O-D pair can then be represented as a link between the POI Topic distributions. With regards to the directions and the time slots of the trip, we propose to collaboratively generate the trip time from the topic-wise specific distribution, which constitutes a *POI Link*. The trip purpose can then be explained semantically from the inferred POI Links. In this way, we jointly model the inter-correlation among origins, destinations, and the trip time. With the learned model, we can uncover the POI Topics for both the origins and destinations to explain the mixed functionality of the origins and destinations, and moreover, we can identify the semantic meanings of trip purposes for each trajectory with the inter-correlations between the POI Topics of the O-D pairs and the trip time. Finally, we present extensive experimental results to demonstrate the effectiveness of our model for spotting trip purposes with interpretation and explanations. Furthermore, quantitative analysis in predicting the destinations and trip time is conducted to validate the predictive power of the proposed model.

2 PRELIMINARIES

Trip purposes reflect the possible activities of mobile users at destinations. Identifying the trip purposes can help discover interesting human mobility patterns and further uncover the volume of traffic flows between different functional regions in a city. For example, if we can identify the trip purpose from the massive traffic flows such as going to work, we can better estimate the capacity of specific traffic flows (e.g., from residential areas to Central Business District in the early morning), which, moreover, can enhance smart urban planning.

Nowadays, most taxicabs are installed with GPS devices, and the latitude and longitude of the pickup point (i.e., the origins (O)) and the dropoff point (i.e., destination (D)) of a trip are recorded, namely O-D pair, along with the time period of the trip. However, such O-D pairs can only reveal the directions of taxi trips, with no semantic meanings to reveal the purpose of the trips.

Table 1. Activity Types with Related POIs

Activity Types	Related POIs
In-home	Residential POIs
Work-related	Companies, governments, institutions, etc.
Transportation transfer	Airports, railway stations, bus stops, subway stations, etc.
Dinning	Bars, restaurants, beverages, etc.
Shopping	Shopping malls, supermarkets, stores, etc.
Recreation	Museums, libraries, parks, movie theaters, etc.
Schooling	University, Primary, middle and high schools, kindergartens, etc.
Lodging	Hotels, motels, etc.
Medical	Hospitals

As a matter of fact, the neighborhood environment at both ends of a trip can help explain the activities and corresponding trip purposes, while the neighborhood environment of a specific point in the city can be depicted by the nearby Point of Interest (POI). As can be seen from Table 1, different POIs are related to certain activity types. Therefore, by augmenting the O-D pairs with neighborhood environment, the information of the trips is greatly enriched and the trip purposed can be inferred. For example, if the origin place is *Newport, New Jersey* with residential areas as the neighborhood environment, while the destination is *Wall Street* at lower Manhattan in New York City, where the place is surrounded by POIs such as office buildings, the trip can then be identified as working purposes. Therefore, we define a *POI Augmented O-D pair* to better uncover the trip purposes.

Definition 1 (Augmented O-D Pair). A POI augmented O-D pair refers to the O-D pair that is augmented by various categories of POIs in the neighborhood of the origin and destination points (within a predefined radius, e.g., 200 meters).

In reality, there are a variety of POIs near a point in the city, since a place is usually blended with multiple functionality. For example, a residential area can be mixed with dining restaurant, parks, and so on, and as illustrated in previous studies (Fu et al. 2015), a residential area with high investment values should be mixed with different types of functionality, that is, balanced categories of POIs. Therefore, the activity types of a given point cannot be precisely revealed by simply listing all the nearby POIs. In this article, we introduce a latent factor, namely *POI Topic*, to reflect the overall distribution of the neighborhood POIs, and to further uncover the various activity types.

Definition 2 (POI Topic). POI Topic of a point refers to the mixture of POIs in the neighborhood environment of a place in the city, which is analogous to the topic of a document when the predefined neighborhood (e.g., in a radius of 200 meters of the point) is treated as a document while the POIs in the neighborhood are treated as words.

For example, if a place has residential houses, supermarkets, outdoor parks in the neighborhood, the POI Topic of the neighborhood environment can be identified as "home" mixed with "recreation"; while if a point is composed of office buildings, shopping malls, and cinemas nearby, it can be regarded as a "business" and "entertainment" topic. By introducing the latent variable *POI topics*, the augmented O-D pairs can then be represented as linked pairs of topics to further uncover the activity types of the trips. Meanwhile, the time of the taxi trips is also recorded along with the augmented O-D pairs, thus the *POI Link* can be defined as follows.

Definition 3 (POI Link). A POI link is defined as a linked pair of POI topics, along with the trip time, which can be represented by a three element tuple (**POI Topic**$_{origin}$, **Trip Time**, **POI Topic**$_{destination}$) to reveal the trip purpose in semantic level.

We use an example to explain the definition of *POI Link*. if a trip happens at 7:00 a.m., the origin place is annotated with a topic 'Residential', and the destination is annotated with topic "Workplace," the corresponding POI link for the model can be represented with a tuple ("Residential," 7:00 a.m., "Workplace"). Then, the POI link can be semantically explained as working oriented trips.

In a nutshell, by introducing the previous concepts, the trip purpose can then be discovered by first augmenting the O-D pairs with neighborhood POIs, and then the semantic meanings of a trip can possibly be discerned from the POI Link that is essentially the POI topics at the two ends of the trip, along with the time period of the trip. More formally, the problem of identifying trip purpose of a taxi trajectory can be defined as follows.

Definition 4 (Problem Statement). Given the GPS trajectories of taxicabs in a city, which include the O-D pairs and the time periods of the taxi trips, as well as the surrounding POIs of origin and destination regions, the goal of the problem is to infer the trip purposes of the trajectories.

According to the previous definitions and the problem statement, in this article, each taxi GPS trajectory is simplified as an O-D pair that has a pickup point, a dropoff point, as well as the corresponding time periods of the trip (we assume that the trip is inside the city and usually less then 1h, thus we only retain the hour of the day for a trip). Essentially, the task of the problem can be decomposed as follows: (i) augment the O-D pairs with the neighborhood POIs; (ii) infer the POI topic of the origins and destinations; (iii) jointly model the POI topics and the time periods of the trip to identify different types of POI links, to reveal the latent semantic meanings of trip purposes.

3 POI LINK MODEL FOR SPOTTING TRIP PURPOSES

We present our proposed model for spotting trip purposes. Specifically, we first introduce the general idea, then develop a generative model, and finally provide the parameter estimation approach.

3.1 General Ideas

Following the definition of *Augmented O-D pairs*, each point in the city is augmented by the neighborhood POIs. Therefore, the points can be regarded as a mixture over the neighborhood POIs and then the blended functionality of the neighborhood can be further illustrated by the introduced latent factor *POI Topics*. The points in the city include the origins and destinations of taxi trips, which can be regarded as documents, while the POIs can be regarded as words. Similar to topic model such as LDA (Blei et al. 2003) that cluster co-occurred words in a topic with particular semantic meanings, the POI topic can reveal the POIs that always show up together in certain places, indicating specific functionality, which corresponds to particular activity types.

According to the definition of *POI Link*, the purpose of a trip can be represented as linkage between the POI Topics of both the origins and destinations, since the activity types can be revealed by the distribution of POI Topics, and moreover, the trip time of the POI link can further indicate the purpose of the trip between regions of different functionality. For example, a trip from the "Home"-related POI Topic to "Work"-related POI Topic can be naturally inferred as for working purpose, while if the trip happens in the early morning, the inference for the purpose would be more accurate.

However, both the origins and destinations are mixtures of POI Topics, and it remains to be a challenging task to link the POI Topics of the origins and the destinations. As a matter of fact, many

efforts have been made to establish connections between different objects. For example, Mixed Membership Stochastic Block (MMSB) model (Airoldi et al. 2008) is originally proposed to model the links between entities, where a Bernoulli distribution is introduced to capture the existence of links between pairwise topics. Several work has adapted the ideas in modeling the relationship between documents by combining MMSB with LDA, in which a Bernoulli distribution is defined over the pairwise topics to model whether they have a connection. However, in this article, we aim to model not only the links between the origin and destinations, but also the particular time periods of the links, thus we introduce a time distribution over pairwise topics to generate the actual trip time.

3.2 Generative Process

In previous sections, the trip purpose can be represented as *POI Links*. Therefore, we propose to model the generative process of *POI Links*. To begin with, the O-D pairs are augmented by the neighborhood POIs, in analogy to LDA model, the neighborhood environment around the point can be treated as a document, with each POI falling in a predefined circle as words, and the POIs can be generated in a probabilistic way as words in LDA. To be more specific, each point in the city can be represented as a multinomial distribution θ over POI topics, which is generated from a Dirichlet prior α. Then, for each POI in the neighborhood of a given point, a POI topic is first generated from the distribution θ, and the POI is generated from the topic-specific POI distribution ψ. Since the pickup points and dropoff points may not always share the same space in the city, or in other words, the origin and destination points may not overlap geographically, the Augmented O-D pairs are modeled, respectively, as separate documents.

In modeling the POI link of the taxi trips, we exploit similar probabilistic generative process as in MMSB, which is, the latent POI topics are first generated for the origin and destination separately, and then a link is generated from the pairwise POI topics. With regards to the definition of *POI Link*, except for whether the link exists between two points, the POI link also contains the time information of the trip. In this case, we exploit a pairwise topic specific multinomial distribution over time slots, that is, $\xi^t_{z_o,z_d}$, showing the probability that time t belongs to the pairwise topic z_o and z_d, which is generated from a Dirichlet prior η. Specifically, for each O-D pair, the POI topics z_{od} and z_{do} are generated from the multinomial distribution over topics θ, respectively, and then the pairwise topics together generate the time slots for the trip from the multinomial distribution over time specified by the topic pairs ξ, revealing how the POI distribution of the pickup and dropoff point interact with different time periods. More formally, the related mathematical notations are listed in Table 2, and we propose the POI Link model (PLM) with the generative process shown in Table 3.

3.3 Model Inference

First, according to the generative process of the PLM, the joint probability distribution of the model is

$$\begin{aligned} &p(\mathbf{w}, \mathbf{t}_{od}, \mathbf{z}_o, \mathbf{z}_d, \mathbf{z}_{od}, \mathbf{z}_{do}, \boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\zeta}; \boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= p(\mathbf{w}|\mathbf{z}_o; \boldsymbol{\varphi}) p(\mathbf{t}_{od}|\mathbf{z}_{od}, \mathbf{z}_{do}, \boldsymbol{\zeta}) p(\mathbf{z}_o|\boldsymbol{\theta}) p(\mathbf{z}_d|\boldsymbol{\theta}) \\ &\quad p(\mathbf{z}_{od}|\boldsymbol{\theta}) p(\mathbf{z}_{do}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) p(\boldsymbol{\varphi}|\boldsymbol{\beta}) p(\boldsymbol{\zeta}|\boldsymbol{\eta}) \end{aligned} \tag{1}$$

In Gibbs Sampling, we should sample the latent assignments for the POI Topics as in Figure 1. Overall, there are four different latent assignments of POI topics in the model, including the latent assignments for the POI topics of the origin and destination points, as well as the POI Topics for the pairwise points in generating the POI Links. Since the pairwise topics are generated in a similar probabilistic process, the full conditionals for the POI topics are also similar. To note that all the

Table 2. Symbol Notations

Symbol	Definition
$\boldsymbol{m} = \mathbf{o} \cup \mathbf{d}$	The set of points in the city {1, . . . , m, . . . , }, which is the union of the origins **o** and destinations **d**.
$\mathbf{w}$	The set of POIs in the city
$\mathbf{z}_o$, $\mathbf{z}_d$	POI Topics for augmented origin point o and destination point d
$\mathbf{z}_{od}$, $\mathbf{z}_{do}$	POI topics for an O-D pair <o,d>, respectively
N_o, N_d	Number of POIs near point o and d
$\boldsymbol{\theta}$	The distribution of POI topics over points in the city
$\boldsymbol{\varphi}$	The distribution of POIs over POI topics
$\boldsymbol{\zeta}$	The distribution of time periods over topic pairs
$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\eta}$	The Dirichlet prior for $\boldsymbol{\theta}$, $\boldsymbol{\varphi}$, $\boldsymbol{\zeta}$
M	Total number of points of the taxi trajectories including the origins and destinations
M_{od}	Total number of O-D pairs
V	Total number of POI categories
K	Number of POI topics
T	Number of time buckets

Table 3. The Generative Process of the POI Link Model

For each origin point $o \in \{1, \ldots, M_o\}$
  Generate the distribution of POI topics of origin point: $\theta_o \sim \text{Dir}(\boldsymbol{\alpha})$
  Generate the distribution of POIs over POI topics: $\varphi_k \sim \text{Dir}(\boldsymbol{\varphi}|\boldsymbol{\beta})$
  For each POI $i \in \{1, \ldots, \boldsymbol{N}_o\}$
    Generate POI topic for the POI i: $\mathrm{z}_{oi} \sim \text{Multi}(\boldsymbol{\theta})$
    Generate each POI i falling in the circle of the origin o: $\mathrm{w}_{oi} \sim \text{Multi}(\boldsymbol{\beta})$
For each destination point $d \in \{1, \ldots, M_d\}$:
  Generate the distribution of POI topics of destination: $\theta_d \sim \text{Dir}(\boldsymbol{\alpha})$
  Generate the distribution of POIs over POI topics: $\varphi_{k'} \sim \text{Dir}(\boldsymbol{\varphi}|\boldsymbol{\beta})$
  For each POI $i \in \{1, \ldots, \boldsymbol{N}_d\}$
    Generate POI topic for the POI i: $\mathrm{z}_{di} \sim \text{Multi}(\boldsymbol{\theta})$
    Generate each POI i falling in the circle of the destination d: $\mathrm{w}_{di} \sim \text{Multi}(\boldsymbol{\beta})$
For each O-D pair $j = (o, d)$:
  Generate the distribution of time periods over topic pairs $\boldsymbol{\zeta} \sim \text{Dir}(\boldsymbol{\eta})$
  Generate POI topics for the jth O-D pair: $\mathrm{z}_{od,j} \sim \text{Multi}(\boldsymbol{\theta}_o)$, $\mathrm{z}_{do,j} \sim \text{Multi}(\boldsymbol{\theta}_d)$
  Generate time slot for the jth O-D pair: $\mathrm{t}_j \sim \text{Multi}(\boldsymbol{\zeta})$

latent POI topics are generated from the multinomial distribution $\boldsymbol{\theta}$, sharing the same Dirichlet prior $\boldsymbol{\alpha}$, thus we can obtain the collapsed from by integrating out $\boldsymbol{\theta}$ as follows:

$$\begin{aligned} & p(\mathbf{z}_o, \mathbf{z}_{od}, \mathbf{z}_d, \mathbf{z}_{do}|\boldsymbol{\alpha}) \\ &= \int p(\mathbf{z}_o|\boldsymbol{\theta}) p(\mathbf{z}_d|\boldsymbol{\theta}) p(\mathbf{z}_{od}|\boldsymbol{\theta}) p(\mathbf{z}_{do}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\theta} \\ &= \prod_{m=1}^{M} \frac{\Delta(\vec{n}_o^m + \vec{n}_d^m + \vec{n}_{od}^m + \vec{n}_{do}^m + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})}, \end{aligned} \tag{2}$$

where $\vec{n}_o^m = \{n_o^{m,k}\}_1^K$ represents the number of origin points assigned to the POI topic k when the points are positioned at m, and $\vec{n}_{od}^m = \{n_{od}^{m,k}\}_1^K$ represents the number of OD pairs with the

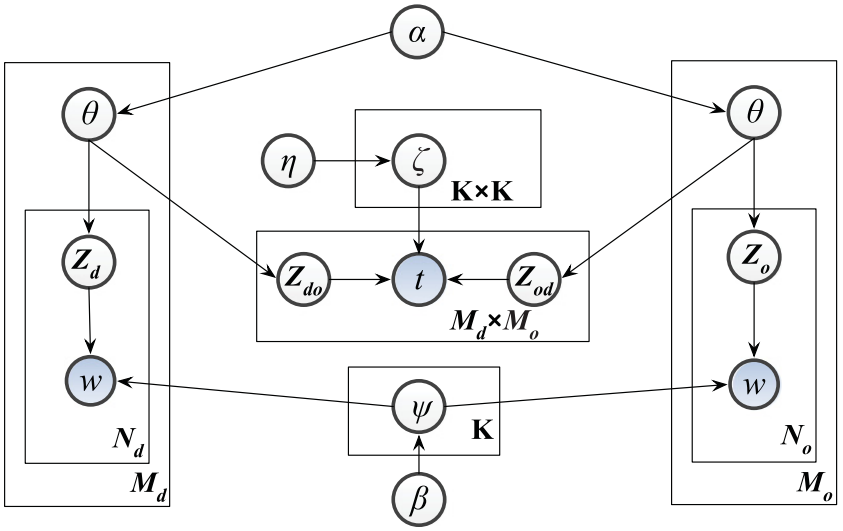


Fig. 1. POI Link Model.

origin points assigned to the POI topic k when they are positioned at m, while the notation $\vec{n}^m_{do}$ are similar for destination points. Meanwhile, the Dirichlet Delta function can be extended as $\Delta(\vec{x}) = \frac{\prod_{k=1}^{|\vec{x}|}\Gamma(x_k)}{\Gamma(\sum_{k=1}^{|\vec{x}|} x_k)}$.

Then, we can derive the latent assignment of POI topic $\mathrm{z}_{o,i}$ for a POI i near the origin point o at position m:

$$\begin{aligned}
&p(\mathrm{z}_{oi} = k|\mathbf{w}, \mathbf{t}_{od}, \mathbf{z}_{o,\neg i}, \mathbf{z}_d, \mathbf{z}_{od}, \mathbf{z}_{do}) \\
&= \frac{p(\mathbf{w}, \mathbf{t}_{od}, \mathbf{z}_o, \mathbf{z}_d, \mathbf{z}_{od}, \mathbf{z}_{do})}{p(\mathbf{w}_{o,\neg i}, \mathbf{t}_{od}, \mathbf{z}_{o,\neg i}, \mathbf{z}_d, \mathbf{z}_{od}, \mathbf{z}_{do})} \propto \frac{p(\mathbf{w}|\mathbf{z}_o;\boldsymbol{\varphi})}{p(\mathbf{w}_{o,\neg i}|\mathbf{z}_{o,\neg i};\boldsymbol{\varphi})} \frac{p(\mathbf{z}_o|\boldsymbol{\theta})}{p(\mathbf{z}_{o,\neg i}|\boldsymbol{\theta})} \\
&= \frac{\Delta(\vec{n}^{(w)}_k + \boldsymbol{\beta})}{\Delta(\vec{n}^{(w)}_{k,\neg i} + \boldsymbol{\beta})} \frac{\Delta(\vec{n}^m_o + \vec{n}^m_d + \vec{n}^m_{od} + \vec{n}^m_{do} + \boldsymbol{\alpha})}{\Delta(\vec{n}^m_{o,\neg i} + \vec{n}^m_d + \vec{n}^m_{od} + \vec{n}^m_{do} + \boldsymbol{\alpha})} \\
&= \frac{n^{(w)}_{k,\neg i} + \beta_w}{\sum_{w=1}^{V}(n^{(w)}_{k,\neg i} + \beta_w)} \cdot \frac{n^{m,(k)}_{o,\neg i} + n^{m,(k)}_d + n^{m,(k)}_{od} + n^{m,(k)}_{do} + \alpha_k}{\sum_{k=1}^{K} n^{m,(k)}_{o,\neg i} + n^{m,(k)}_d + n^{m,(k)}_{od} + n^{m,(k)}_{do} + \alpha_k},
\end{aligned} \tag{3}$$

where $\vec{n}^{(w)}_k = \{n^{(w),v}_k\}^V_1$ represents the number of POIs assigned to the POI Topic k.

In the same way, the sampling process for the latent assignments of the POI topics z_d for the destination d is the same as z_o:

$$\begin{aligned}
&p(\mathrm{z}_{di} = k|\mathbf{w}, \mathbf{t}_{od}, \mathbf{z}_o, \mathbf{z}_{d,\neg i}, \mathbf{z}_{od}, \mathbf{z}_{do}) \\
&\propto \frac{n^{(w)}_{k,\neg i} + \beta_w}{\sum_{w=1}^{V}(n^{(w)}_{k,\neg i} + \beta_w)} \cdot \frac{n^{l,(k)}_o + n^{l,(k)}_{d,\neg i} + n^{l,(k)}_{od} + n^{l,(k)}_{do} + \alpha_k}{\sum_{k=1}^{K} n^{l,(k)}_o + n^{l,(k)}_{d,\neg i} + n^{l,(k)}_{od} + n^{l,(k)}_{do} + \alpha_k}.
\end{aligned} \tag{4}$$

Moreover, in the proposed model, the pairwise POI topics together generate the trip time, thus the two latent variable $\mathrm{z}_{od,j}$ and $\mathrm{z}_{do,j}$ are sampled simultaneously as follows, the position of the

OD pair is m and l, respectively:

$$
\begin{aligned}
&p(\mathrm{z}_{od,j} = k, \mathrm{z}_{do,j} = k' | \mathbf{w}, \mathbf{t}_{od}, \mathbf{z}_{o,\neg i}, \mathbf{z}_d, \mathbf{z}_{od,\neg j}, \mathbf{z}_{do,\neg j}) \\
&= \frac{p(\mathbf{w}, \mathbf{t}_{od}, \mathbf{z}_o, \mathbf{z}_d, \mathbf{z}_{od}, \mathbf{z}_{do})}{p(\mathbf{w}_o, \mathbf{t}_{od,\neg j}, \mathbf{z}_o, \mathbf{z}_d, \mathbf{z}_{od,\neg j}, \mathbf{z}_{do,\neg j})} \\
&\propto \frac{p(\mathbf{z}_{od} | \mathbf{z}_{od}, \mathbf{z}_{do}; \boldsymbol{\zeta})}{p(\mathbf{t}_{od,\neg j} | \mathbf{z}_{od,\neg j}, \mathbf{z}_{do,\neg j}; \boldsymbol{\zeta})} \frac{p(\mathbf{z}_{od} | \boldsymbol{\theta}) p(\mathbf{z}_{do} | \boldsymbol{\theta})}{p(\mathbf{z}_{od,\neg j} | \boldsymbol{\theta}) p(\mathbf{z}_{do,\neg j} | \boldsymbol{\theta})} \\
&= \frac{\Delta(\vec{n}^{(t)}_{od} + \boldsymbol{\eta})}{\Delta(\vec{n}_{od,\neg j} + \boldsymbol{\eta})} \frac{\Delta(\vec{n}^m_o + \vec{n}^m_d + \vec{n}^m_{od} + \vec{n}^m_{do} + \boldsymbol{\alpha})}{\Delta(\vec{n}^m_o + \vec{n}^m_d + \vec{n}^m_{od} + \vec{n}^m_{do} + \boldsymbol{\alpha})} \frac{\Delta(\vec{n}^l_o + \vec{n}^l_d + \vec{n}^l_{od} + \vec{n}^l_{do} + \boldsymbol{\alpha})}{\Delta(\vec{n}^l_o + \vec{n}^l_d + \vec{n}^l_{od,\neg j} + \vec{n}^l_{do,\neg j} + \boldsymbol{\alpha})} \\
&= \frac{n^{(t)}_{kk',\neg j} + \eta_{kk'}}{\sum_{t=1}^T n^{(t)}_{kk',\neg j} + \eta_{kk'}} \cdot \frac{n^{m,(k)}_o + n^{m,(k)}_d + n^{m,(k)}_{od,\neg j} + n^{m,(k)}_{do,\neg j} + \alpha_k}{\sum_{k=1}^K n^{m,(k)}_o + n^{m,(k)}_d + n^{m,(k)}_{od,\neg j} + n^{m,(k)}_{do,\neg j} + \alpha_k} \\
&\cdot \frac{n^{l,(k)}_o + n^{l,(k)}_d + n^{l,(k)}_{od,\neg j} + n^{l,(k)}_{do,\neg j} + \alpha_k}{\sum_{k=1}^K n^{l,(k)}_o + n^{l,(k)}_d + n^{l,(k)}_{od,\neg j} + n^{l,(k)}_{do,\neg j} + \alpha_k}.
\end{aligned} \tag{5}
$$

Moreover, the posterior estimates for time distribution over pairwise POI Topics ζ, the POI Topic distribution for points θ, and the POI distribution over POI Topics φ can be computed in the following:

$$
\begin{aligned}
\zeta^t_{kk'} &= \frac{n^{(t)}_{kk'} + \eta_{kk'}}{\sum_{t=1}^T n^{(t)}_{kk'} + \eta_{kk'}}, \\
\theta^k_m &= \frac{n^{m,(k)}_o + n^{m,(k)}_d + n^{m,(k)}_{od} + n^{m,(k)}_{do} + \alpha_k}{\sum_{k=1}^K n^{m,(k)}_o + n^{m,(k)}_d + n^{m,(k)}_{od} + n^{m,(k)}_{do} + \alpha_k}, \\
\varphi^w_k &= \frac{n^{(w)}_k + n^{(w)}_{k'} + \beta_w}{\sum_{w=1}^V n^{(v)}_k + n^{(w)}_{k'} + \beta_w}.
\end{aligned} \tag{6}
$$

3.4 Model Applications

3.4.1 Spotting the Trip Purpose. By modeling the pre-defined *POI Links*, we can uncover the latent structures of the taxi trips with regards to the neighborhood POIs of the pickup and dropoff points, the trip time, and the links between the the origins and destinations, which provides clear clues to spot the trip purpose by explaining the semantic meanings. To be specific, each point in the taxi trajectories is represented as a mixture over certain POI Topics as indicated by the inferred parameter $\boldsymbol{\theta}$, thus we then let the POI Topic with the largest probability values as the assignment for the point m, that is,

$$
k^*_m = \underset{k}{\operatorname{argmax}}\ \theta^k_m. \tag{7}
$$

The O-D pairs with the same assigned POI Topic pairs (k^*_o, k^*_d) are clustered into the same type of POI links, and thus, each O-D pair can be categorized into one of the K^2 clusters. In company with the time distribution of the topic pairs, we can further explain the purpose of each taxi trip.

3.4.2 Predicting the Destination. By inferring the parameters from PLM, we can estimate the destination given the origin and the time period of the trip. To be more specific, given the origin point o and the trip time t, we aim to estimate the probability that the destination of the trip is d, that is, $p(d|o,t)$. According to the proposed model PLM, the neighborhood of each point m is

augmented with POIs, thus the origin and destination point can be regarded as documents $\mathbf{w}_o$ and $\mathbf{w}_d$, with the nearby POIs as words. Concretely, we assume all the points and the neighborhood POIs in the city are known, the probability can be computed according to Bayes rule as follows:

$$\begin{aligned} p(\mathbf{w}_d|\mathbf{w}_o, t) &= \frac{p(\mathbf{w}_o, \mathbf{w}_d, t)}{\sum_{d'} p(\mathbf{w}_o, \mathbf{w}_{d'}, t)} \\ &\propto \sum_{\mathrm{z}_{od}\mathrm{z}_{do}} p(t|\mathrm{z}_{od}\mathrm{z}_{do}) p(\mathrm{z}_{od}|\theta_{od}) p(\mathrm{z}_{do}|\theta_{do}) p(\mathrm{w}_d|\theta_d), \end{aligned} \tag{8}$$

where $p(\mathrm{w}_d|\theta_d) = \prod_w \sum_{z_d} p(w|z_d) p(z_d|\theta_d)$.

3.4.3 Predicting the Trip Time. Meanwhile, we can also derive the probability of the time slots of a trip given the origin and the destination by inferring the posterior from PLM. Concretely, given the origin point o, which can be augmented as a bunch of POIs $\mathbf{w}_o$ and the destination point d, which can be augmented as $\mathbf{w}_d$, we aim to estimate the probability that the trip takes place at time t, that is, $p(t|\mathbf{w}_o, \mathbf{w}_d)$. Similarly, we also assume that all the points and the neighborhood POIs in the city are known, and then the probability can be computed according to Bayes rule as follows:

$$\begin{aligned} p(\mathbf{t}|\mathbf{w}_o, \mathbf{w}_{\boldsymbol{d}}) &= \frac{p(t, \mathbf{w}_o, \mathbf{w}_d)}{\sum_{t'} p(t', \mathbf{w}_o, \mathbf{w}_d)} \\ &\propto \sum_{\mathrm{z}_{od}\mathrm{z}_{do}} p(t|\mathrm{z}_{od}\mathrm{z}_{do}) p(\mathrm{z}_{od}|\theta_{od}) p(\mathrm{z}_{do}|\theta_{do}). \end{aligned} \tag{9}$$

4 EXPERIMENT RESULTS

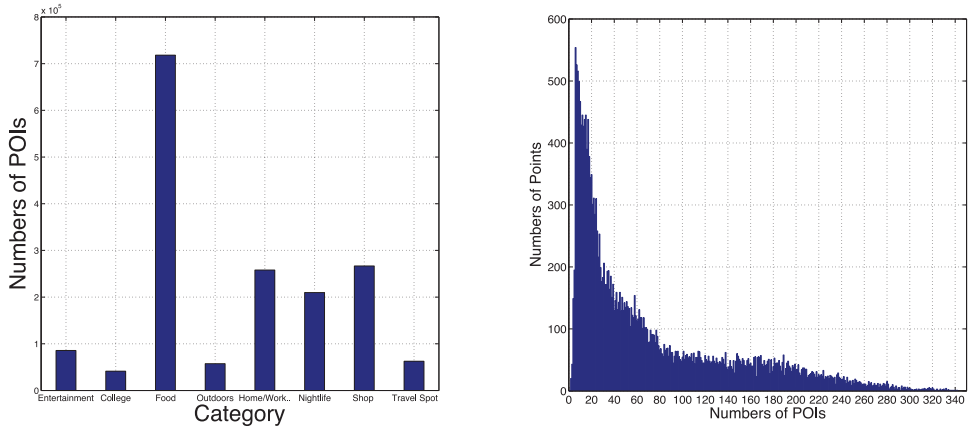
In this section, we first introduce the settings of our experiments and then present the results of the designed studies.

4.1 Data Description

We use the following datasets in New York City to evaluate the proposed POI Link Model.

Human Mobility Data: We collected the GPS trajectory data set generated by NYC taxis from 01/2015 to 06/2015. In the original data set, there are more than 70,000,000 trajectories from about 700,000 pickup points to more than 1,000,000 dropoff points. To uncover the semantic meanings of trip purposes and make the model computable, we sample a more compact dataset in the experiments. Specifically, we filtered out the desolate points that have less than 5 POIs in the neighborhood and the points with few trips (less than five). In addition, we also filtered out the obviously incorrect trajectories and points and the points that do not belong to NYC, which finally result in a dataset with 188,363 trajectories containing 13,455 pickup points and 17,926 dropoff points.

Point of Interest Data and Augmented O-D pairs: We extracted the POI information from the NYC checkin data of Fourquare.com. There are 8 POI categories with top-levels and 241 sub-categories of POIs. With the POI information distributed in the city, we map the POIs to the specific points of the OD pairs of taxi trajectories and generate *Augmented O-D pairs* according to Definition 1, and in total there are 66,275 POIs under the categories. Specifically, a circle of radius 200m is generated for each point, and the POIs falling in the circle are used to augment the points, making augmented O-D pairs. The frequency of each POI category and the distribution of the number of POIs around the points in NYC are presented in Figure 2. We can see from the figure that the category of "food"-related POIs take up the most fraction of all the POIs, and meanwhile most points have less than 100 POIs in the neighborhood.



(a) The POI categories. (b) The distribution of the number of POIs

Fig. 2. POI categories of the dataset.

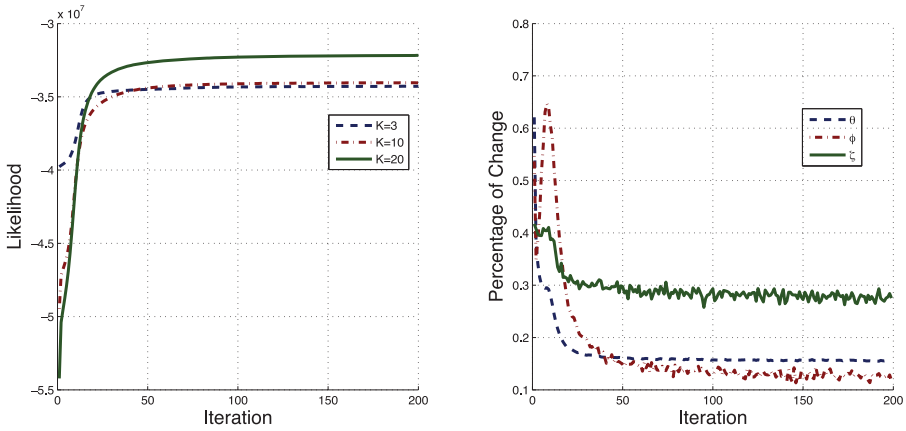


Fig. 3. Log likelihood and percentage of changes in the parameters $\boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\varphi}$ along iterations.

4.2 Model Setup and Model Convergence

In implementing the POI Link Model, we first set the number of POI topics to be 10, and set the hyper-parameters to be symmetric. Meanwhile, we divide the 24h of each day into 12 slots, with two adjacent hours as the same time slots; and we also treat the workdays and weekends differently, such that we have 24 different time slots in total.

Following the Gibbs Sampling process, the parameters are estimated through multiple iterations, and we observe the log likelihood and the changes of percentage in the parameters with the number of iteration. It is shown that the log likelihood converges to a steady value after 100 iterations, and we can also see from Figure 3 that the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ quickly drops to a constant change rate after initial iterations. While the parameter $\boldsymbol{\zeta}$ is indeed the time distribution over the pairwise, which is a three-dimensional matrix with more elements, thus the percentages of changes are greater than the other two parameters when the model is converging.

Table 4. Representative POIs for POI Topics Extracted by PLM

TOPIC 1	prob.	TOPIC 2	prob.	TOPIC 3	prob.	TOPIC 4	prob.	TOPIC 5	prob.
Coffee Shop	0.056	General Entertain	0.174	College Acad. Bldg	0.071	Lounge	0.121	Italian Rest.	0.052
Clothing Store	0.037	Event Space	0.116	University	0.053	Bar	0.106	Home	0.046
High Tech Outlet	0.035	Other Nightlife	0.063	General Univ.	0.052	Hotel	0.099	Coffee Shop	0.046
Bakery	0.034	Performing Arts	0.050	College Admin. Bldg	0.044	Hotel Bar	0.062	Pizza Place	0.046
Cosmetics Shop	0.034	Music Venue	0.046	Library	0.044	Cocktail Bar	0.055	Sushi Rest.	0.035
Office	0.032	Building	0.030	College Library	0.040	American Rest.	0.047	Barbershop	0.032
Miscellaneous Shop	0.030	General Travel	0.028	College Classroom	0.037	Nightclub	0.033	Bakery	0.031
Gift Shop	0.029	Other Outdoors	0.027	Cafe	0.036	French Rest.	0.028	Japanese Rest.	0.030
Sandwich Place	0.029	Nightclub	0.025	Coffee Shop	0.034	Event Space	0.028	Spa	0.027
Train Station	0.029	Plaza	0.024	College Cafeteria	0.029	Breakfast Spot	0.024	Cosmetics Shop	0.027
TOPIC 6	**prob.**	**TOPIC 7**	**prob.**	**TOPIC 8**	**prob.**	**TOPIC 9**	**prob.**	**TOPIC 10**	**prob.**
Art Gallery	0.121	Bar	0.116	Office	0.176	Bar	0.189	Home	0.209
Chinese Rest.	0.068	Wine Bar	0.061	American Rest.	0.068	American Rest.	0.087	Building	0.098
Clothing Store	0.065	Italian Rest.	0.043	Sandwich Place	0.050	Sports Bar	0.064	Park	0.090
Caf	0.044	Lounge	0.040	Coffee Shop	0.043	Pub	0.049	Other Outdoors	0.065
Scenic Lookout	0.033	Breakfast Spot	0.038	Food Truck	0.037	Burger Joint	0.046	Playground	0.055
Design Studio	0.032	Cocktail Bar	0.034	Cafe	0.037	Home	0.036	Speakeasy	0.029
Museum	0.029	Coffee Shop	0.034	Bar	0.036	Gay Bar	0.025	Strip Club	0.030
Bakery	0.027	Cafe	0.033	Deli or Bodega	0.035	Dive Bar	0.024	Laundromat	0.024
Office	0.026	American Rest.	0.033	Building	0.034	Karaoke Bar	0.024	Scenic Lookout	0.022
Dessert Shop	0.024	Home	0.027	General Entertain	0.024	Pizza Place	0.023	Gas Station	0.018

4.3 Discovering POI Topics

In the proposed POI Link Model, the neighborhood POIs are augmented for the O-D pairs. Thus, each point in the trajectories is treated as a mixture of POI topics to describe its neighborhood environment. The POI Link Model infers the POI topic distribution of the points, as well as the the distribution of POIs over a specific POI topic. To uncover the semantic meanings of the POI topics, we analyze the parameters φ_k^w, and list the most representative POIs for each POI topic in Table 4.

We see from Table 4 that each extracted POI topic is a mixture of several representative POIs with similar functionality, corresponding to a similar type of activities. For example, Topic 1 is represented with different kinds of shops, which can be explained as "Shopping"-related topic, and we can identify other semantic meanings from the POIs including "Travel & Entertainment" (Topic 2), "University" (Topic 3), "Hotel & Bar" (Topic 4), "Restaurant" (Topic 5), "Arts" (Topic 6), "Bar" (Topic 7 and 9), "Office" (Topic 8), "Home" (Topic 10).

It is notable to find that in POI Topic 8, "Office" takes up the highest probability, while the topic also contains some other food-related POIs such as "Sandwich Place," "Food Truck," and so on. This exactly matches the environment of working place where the working people buy simple foods and coffee for lunch. We also find that Topics 4, 7, and 9 are all bar-related, but there exist tremendous differences by taking a closer look at the composition of such topics. In Topic 4, except for "Bar," we can also see POIs such as "hotel," "hotel bar" with high probability, indicating that Topic 4 is more like a "hotel"-related topic, with bars, restaurants nearby. While in Topic 9, the representative POIs are mostly various types of bars, showing the topic is a "nightlife"-related topic. It is interesting to note that Topic 1 has both shopping-related POIs and train stations. This may be due to the fact that the center business districts can also be the traffic centers, which have train stations, subways nearby. Though shops and train stations seem unrelated semantically and do not belong to the same category, they are not in conflict with urban functionality and corresponding trip purposes considered, since it is natural that the shops should be blended with traffic facilities for consumers' convenience.

As a matter of fact, we can directly apply LDA model on all the points and extract topics by treating each point as a document and the nearby POIs as words, without considering the links

Table 5. Representative POIs for POI Topics Extracted by LDA

TOPIC 1	**prob.**	**TOPIC 2**	**prob.**	**TOPIC 3**	**pro.**	**TOPIC 4**	**prob.**	**TOPIC 5**	**prob.**
College Acad. Bldg	0.063	Clothing Store	0.064	Office	0.159	Bar	0.125	Lounge	0.111
University	0.047	Cafe	0.052	Sandwich Place	0.057	Wine Bar	0.048	Bar	0.102
General Univ.	0.046	Chinese Rest.	0.051	Coffee Shop	0.050	Lounge	0.037	Hotel	0.081
College Admin. Bldg	0.040	Coffee Shop	0.046	American Rest.	0.048	American Rest.	0.036	Cocktail Bar	0.059
College Library	0.034	Art Gallery	0.046	Cafe	0.038	Italian Rest.	0.034	Hotel Bar	0.053
Coffee Shop	0.034	Bakery	0.040	Food Truck	0.038	Breakfast Spot	0.028	American Rest.	0.050
Library	0.033	Italian Rest.	0.032	Deli or Bodega	0.035	Home	0.027	Nightclub	0.033
Cafe	0.033	Salon or Barbershop	0.028	Building	0.035	Cocktail Bar	0.024	Italian Rest.	0.032
College Classroom	0.032	Cosmetics Shop	0.025	Bar	0.029	Coffee Shop	0.023	French Rest.	0.031
College Cafeteria	0.026	Miscellaneous Shop	0.024	Rest.	0.023	Cafe	0.023	Breakfast Spot	0.028
TOPIC 6	**prob.**	**TOPIC 7**	**prob.**	**TOPIC 8**	**pro.**	**TOPIC 9**	**prob.**	**TOPIC 10**	**prob.**
General Entertain	0.099	Home	0.108	Event Space	0.100	Home	0.174	Bar	0.103
Office	0.077	Pizza Place	0.052	Art Gallery	0.100	Park	0.115	American Rest.	0.080
American Rest.	0.047	Coffee Shop	0.040	General Entertain	0.080	Building	0.086	Sports Bar	0.040
Plaza	0.041	Italian Rest.	0.038	Scenic Lookout	0.052	Playground	0.080	Burger Joint	0.039
General Travel	0.04	Building	0.036	Other Nightlife	0.043	Other Outdoors	0.066	Italian Rest.	0.03
Other Outdoors	0.039	Deli or Bodega	0.033	Office	0.039	Strip Club	0.032	Pizza Place	0.030
Performing Arts	0.036	Cosmetics Shop	0.023	Nightclub	0.034	Speakeasy	0.027	Arcade	0.028
Building	0.033	Diner	0.023	Music Venue	0.031	Gas Station	0.027	Coffee Shop	0.027
Event Space	0.028	Bakery	0.023	Building	0.028	Laundry	0.023	Mexican Rest.	0.022
Bar	0.027	Salon or Barbershop	0.022	Bar	0.024	Parking Garage	0.023	Bakery	0.022

between points. We can also represent the semantic meanings of each topic with several POIs. We find that though some related POIs can also be clustered in one particular topics, the overall distribution of POIs over topics extracted from LDA are not as clear as that in PLM, and there exist some ambiguous topics. For example, as shown in Table 5, Topic 2 is similar to the mixture of Topic 1 and Topic 6 in Table 4; however, it does not explicitly distinguish "Shopping," "Arts," and "Restaurants"; Topic 6 has "General Entertainment," which is similar to Topic 2 from PLM, while we find that the topic contains other unrelated POIs such as "Office," "Bar," making it hard to explain. This is because LDA model only takes the co-occurrence of POIs in the same region into account, making the topics more ambiguous. We see that Topic 6 extracted by LDA is mixed with "General Entertainment" and "Office," such co-occurrence of POIs can be discovered in some regions such as *Wall Street* and *Time Square*, which are both scenic spots and working places, making a mixed POI topics. However, a working oriented trip and a travel oriented trip may not usually start from the same origins with different trip time, which is not modeled by LDA while exactly modeled PLM. Thus, trips with different purposes can be distinguished, resulting in more coherent and explainable POI Topics extracted by PLM, for instance, Topics 2 and 8 in PLM are well separated in terms of entertainment and working.

In addition, to better illustrate the performance in distinguishing different POI topics, we first annotate each point with one topic, that is, the topic with highest probability. Then, we can compute the average KL-divergence between points with different annotated topics according to the Equation (10). This metric measures how well the model can separate different topics, which is adapted from prior work in validating the degree of distinction of topics (Wang and Mccallum 2006), and the higher values represent better performance. As shown in Figure 4, the proposed PLM can achieve better performance in distinguishing different topics than LDA model:

$$D_{KL}@Z = \frac{\sum_{z=1}^{Z}\sum_{z'=1}^{Z}\sum_{v=1}^{V} p(z|v)log(\frac{p(z|v)}{p(z'|v)})}{(Z*(Z-1))/2}. \tag{10}$$

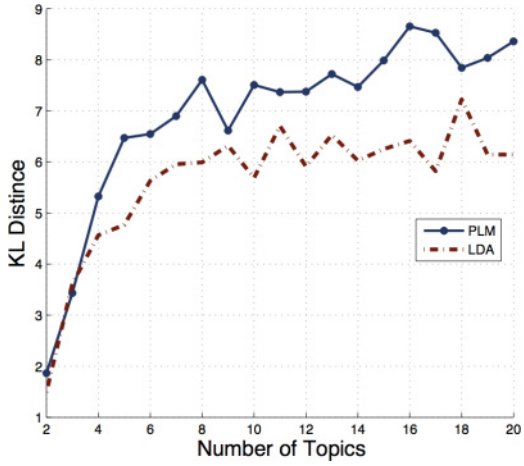
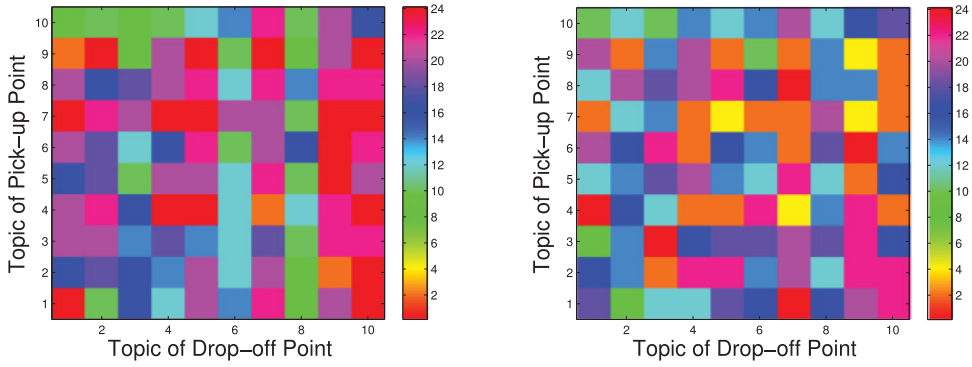


Fig. 4. The KL-Divergence between different topics.



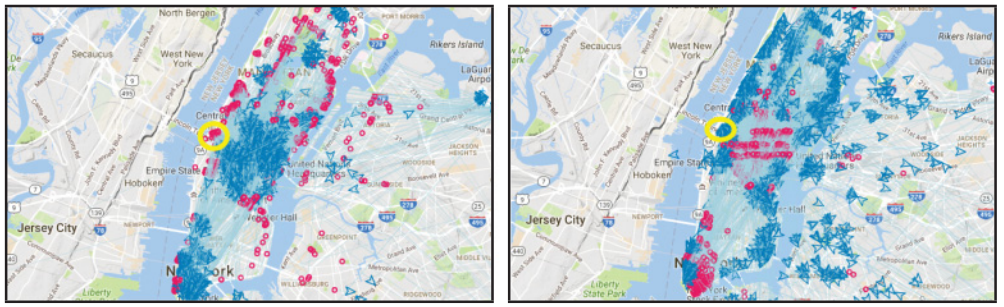
(a) Workday (b) Weekend

Fig. 5. The time distribution of topic pairs.

4.4 Trip Time Distribution for Topic Pairs

In the proposed model PLM, each topic pair is associated with a trip time distribution that can be obtained from the estimated parameters ζ. Concretely, we show the time distribution of each topic pair in Figure 5, with different colors denoting different time periods, and the weekdays and weekends are presented separately.

We can see from that the time distribution for the points of workdays and weekends are different, in which the time distribution for points with different POI topics are more organized with particular spatial-temporal patterns, while the distribution for weekends are more chaotic and with no significant patterns. For example, the POI Topic 8, which can be identified as working district from Table 4, mostly concentrate at 8:00 to 9:00 in the morning when the topic is present at the dropoff points as indicated in Figure 5(a). This corresponds with the intuitions that people take taxi to go to work in the early morning. However, referring to Figure 5(b), the time distribution for the dropoff points with "Work" topic is divergent at different time periods, meaning that there are no regular arrivals for the area at weekend times. Also, the trip to bar-related topics (POI Topics 7



(a) Home to Office (POI Topic 10 to 8) (b) Office to Home (POI Topic 8 to 10)

Fig. 6. The working oriented trips. The red dots denote the origin points and blue arrows denote the destination points.

and 9) fall into the late night, while the time for such trips at weekends are more scattered, and we can observe many trips around midnight (2:00–4:00 a.m.).

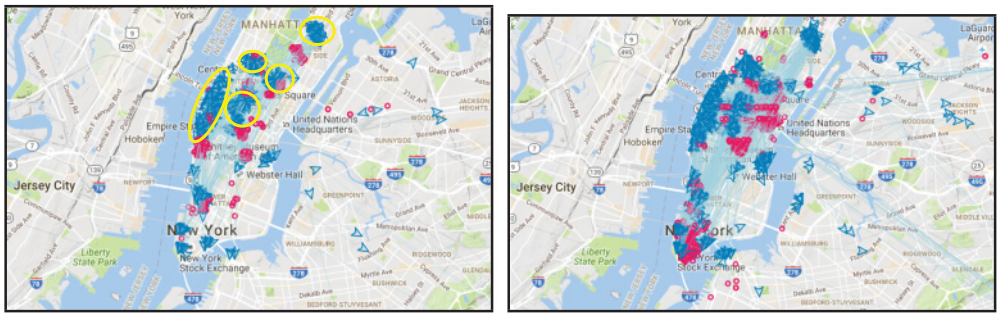
4.5 Annotating Trip Purposes

Since the POI topics are assigned for both ends of the O-D pairs, we can then cluster the O-D pairs according to the latent POI topics. In the taxi trajectories, we know the specific latitude and longitudinal of each point explicitly, we can further visualize each cluster of O-D pairs with particular POI Topic assignments on the map of New York City. Therefore, by identifying the corresponding functionality of the origin and destination in the city map, we can infer the semantic meanings of the POI Link and spot the trip purposes. Considering the extracted POI topics and the particular positions in the map, we show the following identified taxi trajectories and annotate them with the trip purposes separately.

Working oriented trips. It is shown that the POI Topic 8 is related to working sites and Topic 10 is related to home and residential area. Therefore, when the origin points are assigned to "Home" topic and the destination points are assigned to "Office" topic, it is obvious that such trips can be spotted as "working" purpose (Figure 6(a)), and reversely, the trips can be identified as "homing" purpose (Figure 6(b)). By clustering the taxi trajectories according to the POI topics at the two ends, we further visualize the trips in map of New York City. As shown in Figure 6, the points with "working" topic mainly concentrate on two areas, the lower Manhattan (*Wall Street*) and the regions near *Time Square*, which exactly conforms to the main working districts and office buildings; meanwhile, we also observe where points with "Home" topic locate.

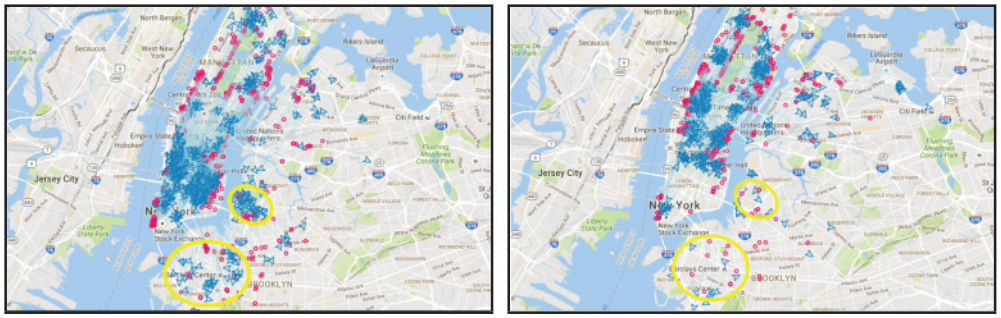
It is interesting to note that many origin points in Figure 6(a) and destination points in Figure 6(b) gather in the point being circled, that is, one end of *Lincoln Tunnel*, which connects NYC with Jersey City. Actually, many people working in NYC choose to live in Jersey City, because the house rentals are less expensive. Furthermore, we can see that the trips with "homing" purpose contain many trajectories from working places to *Queens* and *Brooklyn*, showing the fact that many people work in Manhattan but live farther.

Entertainment oriented trips. POI Topic 2 clusters to the points with "entertainment" POIs nearby. People visit such places for entertainment or travel purposes, and as circled in Figure 7(a), the dropoff points in this cluster matches the famous scenic spots in NYC such as *Empire State*, *The Time Square*, *Lincoln Center*, *The Metropolitan Museum*, and so on. In addition, the pick up points are centered at Topic 4, which mainly consists of hotel-related spots, and moreover, the time of the trips are distributed in all the time of the day. Therefore, such trips may be inferred as the



(a) Hotel to Entertainment (b) Office to Entertainment

Fig. 7. The entertainment oriented trips.



(a) Home to Bar (`POI Topic 7`) (b) Home to Bar (`POI Topic 9`)

Fig. 8. The nightlife oriented trips.

travelers' trips in NYC. Meanwhile, we also find that there are also many trips with "office" topic as origins and "entertainment" as destinations. This may be explained that the working districts are also mixed with many tourists attractions, for example, *Wall Street* is the center of world finance, making it one of the attraction for travelers. Meanwhile, travelers may have to take ships to *Statue of Liberty* by way of Lower Manhattan.

Nightlife oriented trips. The nightlife of NYC is abundant, and we see several POI topics related to different types of bars and restaurants (Topics 7 and 9) with the arrival time at nights as revealed in Figure 5. We analyze the trips from "home"-related topics to "bar"-related topics and such trips can be naturally explained as "hanging out at night". Though both POI Topics 7 and 9 are mixtures of bars and restaurants, they are actually with different types and regions. The visited bars with Topic 7 mainly gather in *Lower Manhattan*, while the bars of Topic 9 are located in *Midtown*. As a matter of fact, we see from Figure 5 that the trip time for Topic 9-related points are even later than POI topics, showing that the bars of Topic 9 open all night and are appealing to the people stay up late. Notably, in Figure 8(a), some "nightlife" trips are located in *Brooklyn*, but such trips merely show up in Figure 8(b), revealing the fact that people are not used to hang out too late at night in that area.

One of the assumptions in this model is that the trips between regions with similar POI distribution at the same time may share similar purposes. Therefore, in an ideal cluster of annotated trips, the trip time should be close or with similar distributions. Therefore, to validate the coherence of the annotated POI Link, we then compute the average time differences of the trips in the same annotated cluster, and the results in comparison with LDA are shown in Figure 9. We see that the

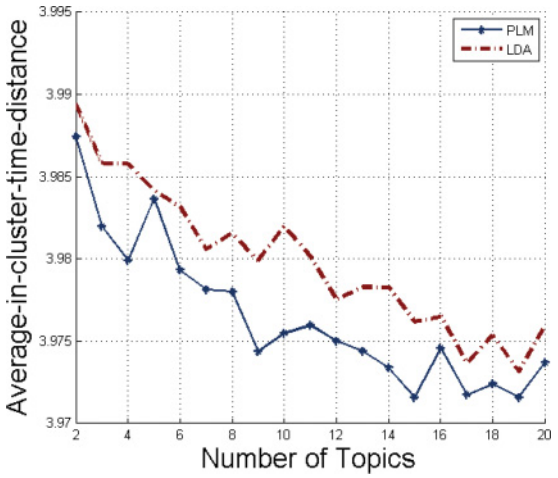


Fig. 9. The average time differences in cluster of trips with annotated purpose.

time differences in the annotated trips inferred PLM are smaller than that from LDA, indicating that the proposed model can identify a more coherent cluster of POI Links.

4.6 Predicting the Destination

In this experiment, we held out a group of trajectories between the origin and destination points as testing set for prediction. According to Equation (8), we can predict the probability of the points to be the destination given the origin and the trip time, and then we can rank all the candidate points and evaluate the prediction accuracy. Specifically, we exploited 150,690 trajectories as training set, and the remaining 37,673 trajectories were held out as the testing set. To better validate the predictive power of the model, for each origin point in the training set, we regard all the destinations that have ever shown up in the training set as the candidate destinations, and we predict the destination for the origin points with more than 10 and less than 200 candidate destinations.

Evaluation Metrics. For each trajectory, there is only one destination given an origin and the trip time, while generally we have a list of ranked K candidate destinations inferred from the model. Therefore, we first use Hit Ratio@K to measure the accuracy of destination prediction, which can be computed as follows:

$$\text{HR@}K = \frac{\sum_{n_o=1}^{N_{tr}} \sum_{n_d=1}^{K} \mathbb{I}(n_o, n_d)}{N_{tr}}, \tag{11}$$

where N_{tr} is the number of trajectories in testing set, and $\mathbb{I}(\cdot)$ is an indicator function. When the n_d^{th} predicted destination is actually the real destination of the origin point n_o, $\mathbb{I}(n_o, n_d) = 1$, and we call the predicted destinations hit the ground truth.

Moreover, since our model captures the links between pairwise POI Topics, we aim to validate whether the neighborhood environment of the predicted points matches the real destinations. Thus, we proposed to compare the POI Topic distribution of the predicted points with that of the real destinations, and we further exploited D_{KL}, that is, the average KL-Divergence of the K predicted points with the real destinations, to measure such similarity:

$$D_{KL}\text{@}K = \frac{\sum_{n_d=1}^{N_{tr}} \sum_{n_d{}'=1}^{K} \sum_{z_d=1}^{K^z} p(z_d|\theta_{n_d}) log(\frac{p(z_d|\theta_{n_d})}{p(z_d|\theta_{n_d{}'})})}{K * N_{tr}}. \tag{12}$$

Table 6. Results of the Destination Predictions

	Precision					KL distance				
Number	PLM	SVD	Frequency	Random	POISIM	PLM	SVD	Frequency	Random	POISIM
1	0.02974	0.02368	**0.03548**	0.00010	0.03483	2.30033	3.61696	**2.22656**	3.04524	3.20850
2	**0.05949**	0.04700	0.03884	0.00016	0.03653	**2.34612**	3.48751	2.69189	3.06278	3.27098
3	**0.08985**	0.07384	0.04077	0.00020	0.03783	**2.38536**	3.37394	2.87709	3.05686	3.29955
4	**0.11910**	0.09970	0.04246	0.00023	0.03887	**2.42182**	3.30388	2.96121	3.05433	3.31567
5	**0.14761**	0.12621	0.04373	0.00029	0.03972	**2.45180**	3.25867	3.02030	3.05551	3.31610
6	**0.17595**	0.15295	0.04553	0.00029	0.04086	**2.47469**	3.22406	3.04886	3.05317	3.31963
7	**0.20608**	0.18107	0.04863	0.00029	0.04168	**2.49166**	3.19372	3.07036	3.05297	3.32744
8	**0.23484**	0.20719	0.05410	0.00036	0.04249	**2.51171**	3.1634	3.08680	3.05534	3.33257
9	**0.26534**	0.23455	0.06529	0.00042	0.04324	**2.53181**	3.13994	3.09864	3.05886	3.33686
10	**0.29501**	0.26374	0.08463	0.00046	0.04370	**2.55479**	3.12423	3.10231	3.05961	3.33691

Baseline Methods. We compare our proposed approach with the following baseline methods:

- —Random: Randomly choose K points from all the points in the city as destinations.
- —Frequency: Given the origin point, we rank the destination points according to the arrival frequency.
- —SVD: We can construct a matrix between the origin and destination points, with each entry being the frequency that the O-D pairs take place in history, and therefore we can apply Singular Value Decomposition to obtain the lower-rank representations for origin and destination.Then, the candidate destinations can be ranked by computing the sum products of decomposed lower rank factors of origins and destinations.
- —POISIM: We first compute the similarity between the points according to the Euclidean distance of the POI distributions. Then, for each origin point o in the test set, we first find N most similar origin points to o in the training set with regards to the POI distribution similarity. Then, we predict the destinations by finding the Top-K frequent destinations starting from the N similar origins.

Prediction Results. By comparing with the baseline methods on the measure of hit ratios, the prediction results are shown in Table 6. We can see that our approach PLM outperforms other baseline methods in overall.

Meanwhile, we also vary the number of the latent POI Topics to validate the predictive power under different parameter setting. Here, we conduct fivefold cross validation and obtain the average hit ratios. As shown in Figure 10, the HR@K fluctuates with the number of POI Topics in a small range, showing that the number of POI topics does not significantly influence the performance on the destination. Moreover, we also changed proportion of training data and testing data, and as shown in Figure 11, we see that the Hit Ratio remains stable at different percentage of training data, while the KL-Divergence changes at different percentage.

4.7 Predicting the Trip Time

In contrast with predicting the destination, we can also apply the inferred parameters to predict the trip time given the origins and destinations. According to Equation (9), we can compute the posterior probability of each time slot given the observed origins and destinations, and we can further make predictions based on the ranking of the probability. In this experiment, we also held the same testing set mentioned previous, and predict the time slots for all the trajectories in the testing set.

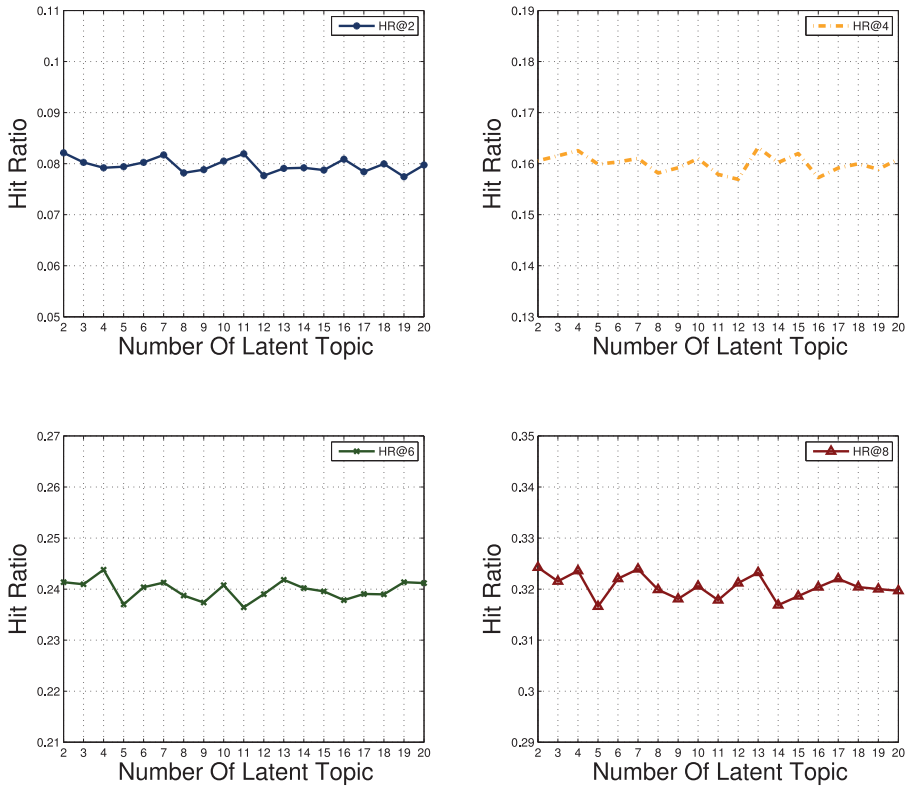


Fig. 10. Hit ratio of destination predictions with different number of POI Topics.

Evaluation Metrics. Similar to the previous experiments in predicting the destinations, we also exploited HR@K to evaluation the prediction results on trip time as follows:

$$\text{HR@}K = \frac{\sum_{tr=1}^{N_{tr}} \sum_{t=1}^{T} \mathbb{I}(tr, t)}{N_{tr}}, \tag{13}$$

where N_{tr} is the number of trajectories in testing set, and $\mathbb{I}(\cdot)$ is an indicator function. When the tth predicted time period is actually the real time bucket of the trace pair tr, $\mathbb{I}(tr, t) = 1$, and we call the predicted time period hit the ground truth.

Baseline Methods. Here are the baseline methods compared with our proposed approach,

— Frequency: Given the origins and the destinations in the training set, we rank the time slots according to the frequency and predict the trip time of the O-D pair with top K time slots.
— KNN: The POI frequency of each category of the O-D pairs are treated as features, and we construct a multi-class classifier with methods K Nearest neighbor, in which the time slots of the O-D pairs are regarded as labels. We then rank the time slots according to the scores output from the classifier model.
— SVM: Similar to KNN, we can construct a multi-class classifiers by SVM, and we can rank the time slots according to the scores for each class.
— Random Forest: We use the random forest method to classify the trajectories into different classes with the time slots as labels. The classified labels output from each tree in the forests

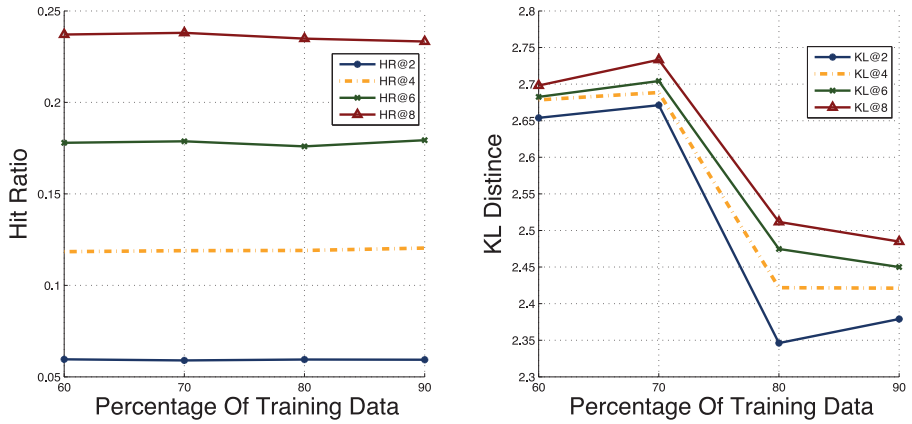


Fig. 11. Hit Ratio of predicted destinations and the average KL-Divergence along with the percentage of training data.

Table 7. Trip Time Prediction Results

	Precision					
Number	PLM	Frequency	KNN	SVM	SVM_LDA	Random Forest
1	**0.114140**	0.033746	0.052478	0.097391	0.110716	0.108937
2	**0.218034**	0.049080	0.111380	0.191596	0.211982	0.202214
3	**0.309612**	0.059990	0.180023	0.272609	0.302631	0.279564
4	**0.385687**	0.018725	0.253895	0.349481	0.377087	0.354153
5	**0.454862**	0.184827	0.324981	0.419637	0.444111	0.420620
6	**0.518116**	0.251002	0.392403	0.489263	0.508030	0.479176
7	**0.575319**	0.320123	0.458259	0.552358	0.566984	0.535609
8	**0.626948**	0.388979	0.523107	0.613437	0.620736	0.588273
9	**0.670454**	0.450615	0.582592	0.663366	0.664986	0.635734
10	**0.706341**	0.538104	0.638760	0.698537	0.702891	0.679452
11	**0.741380**	0.623285	0.685531	0.734319	0.738699	0.718472
12	**0.775197**	0.702201	0.723197	0.766278	0.772383	0.754148

can be regarded as a vote, and we can rank the list of time slots by the outcome of each classifier.

— SVM-LDA: Each point in the city can be represented as a distribution over POIs by applying LDA when we treat the nearby POIs as words and each point as a document. We set the number of topics to be 10, and the probability distribution of the O-D pairs can be used as features, and then we construct an SVM classifier to predict the time slots of each O-D pair.

Prediction Results. The trip time prediction results in comparison with the baselines methods are shown in Table 7. We see that the proposed PLM achieves the best performance among the baseline methods in terms of the Hit Ratio@K, showing the predictive power of the model. To note that SVM-LDA also has good prediction results, which indicates that the fact that the neighborhood environment of the origins and destinations provide clues to predict the trip time; however SVM-LDA only models the POI distribution of the O-D pairs separately, thus it cannot compete with the proposed PLM.

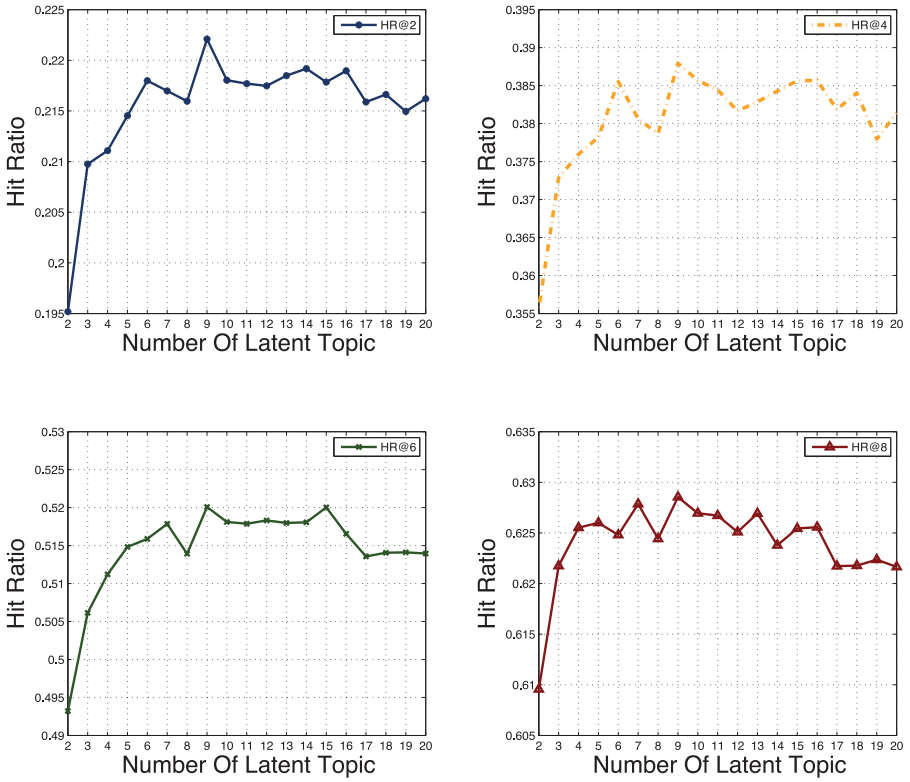


Fig. 12. hit ratio of predicting time periods with different latent topic number K along numbers of recommendations.

We also show the prediction results under different number of POI Topics in Figure 12. We can see that when the prediction results improves along with the increase of the number of POI Topics, and the performance is best when the number reaches 10.

5 CASE STUDY

In this section, we use a case to illustrate the predictive power of the proposed model PLM. As shown in Figure 13, we aim to predict the destinations of a taxicab that picks up passengers near *York Avenue*, between *72th street* and *73th street* at 11:46 p.m. on Tuesday, January 27, 2015, and the POI Topic distribution of the point is shown in the left part of Figure 14, we can see the probabilities of origin point distributed mainly in Topics 5 and 10, therefore we can infer the place as a "Home & Living"-related areas. Along with the destination prediction procedures, we can predict the candidate destinations in terms of the posterior probability $p(\mathbf{w}_d|\mathbf{w}_o, t)$, and rank several destinations according to the probability. In Figure 13, the left part of shows the real trajectory that is presented by a straight line from the origin labeled by "O" to the destination labeled by "D|p1," and the other ten curves show the predicted destinations marked from "p1" to "p10" based the descending order of the calculated probabilities shown in Table 8.

We can see that the predicted destination that ranks at the top exactly corresponds to the actual destinations of the taxi trajectories. Taking a closer look at the neighborhood environment of "p1," it is near the intersection of *E Broadway* and *Clinton Street*. In addition, as shown in the right part of

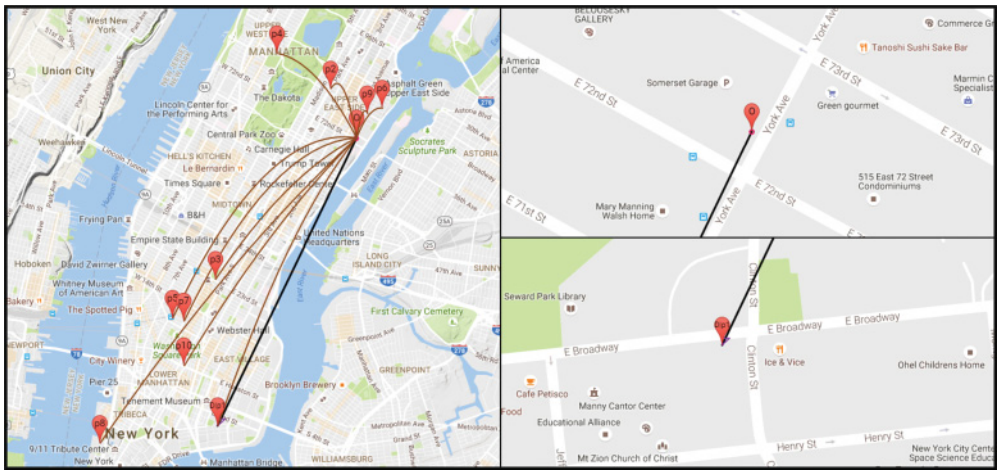


Fig. 13. Predicting the destination origin from *Grand Army Plaza*.

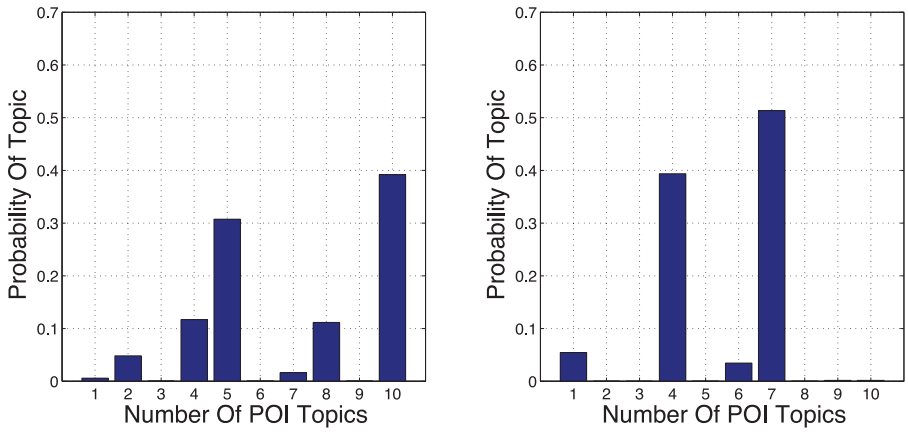


Fig. 14. Distributions of Latent Topics for the OD pair.

Table 8. Predicted Destinations

Predicted Points	Probability	POI Topic
p1	**9.999998e-1**	7
p2	**1.994391e-7**	1
p3	**1.728836e-11**	8
p4	**4.668654e-12**	10
p5	**1.358332e-12**	7
p6	**4.225473e-13**	10
p7	**2.156204e-18**	9
p8	**1.066986e-21**	2
p9	**1.503719e-22**	5
p10	**3.980955e-24**	1

Figure 14, the destination distributed mainly in Topics 4 and 7, showing that the trip is from home to bar for recreation purpose. We can also see from Table 8 that the POI Topics of the predicted destinations concentrate on Topic 7 and 10, which corresponds "Bar"-related and "Home"-related POI Topic, which is deemed reasonable considering the time of the trip is at weekday night. This case shows that the PLM captures the inter-correlations between the neighborhood environment of the O-D pairs and the time of the trip.

6 RELATED WORK

This article is related to the following streams of prior work: (i) inferring trip purpose inference; (ii) mining human mobility data; and (iii) link analysis and topic modeling.

Trip Purpose Inference. With the advent of mobile and GPS technologies, large-scale footprint data are available and have attracted a lot of effort in inferring trip purposes and destination activities using human mobility (Zheng 2015b). The work in Sadilek et al. (2012) incorporated the patterns in friendship formation, the content of peoples messages, and user locations, treated users with known GPS positions as noisy sensors of the location of their friends, and presented a scalable probabilistic model of human mobility for location prediction. The work in Kumar et al. (2015) presented a geographic choice model to consider various factors including distance, rank, and popularity, and ultimately produce an accurate estimate of the likelihood that a user will select one from a set of geographic locations in a region. The work in Lin et al. (2015) proposed a collective iterative classification algorithm to classify the purpose of passengers. The general idea of this work is to first form groups based on the extracted mobility features and collectively classify the purposes of group passengers sharing same destination and same purposes. The work in Zheng et al. (2010b) exploited the location data based on GPS and users' comments at various locations to discover interesting locations and possible activities for recommendations. The work in Furletti et al. (2013) proposed a method to automatically annotate raw trajectories with the activities of users during their movement when tracked by a GPS device. Dewri et al. (2013) proposed a method to determine the destinations of trips using only speed and time data from real world driving trips without accessing GPS traces. Zhu et al. (2014) focused on modeling and inferring the purpose of travel, as well as the activity at the destination of a trip during daily life scenarios. Gao et al. (2012) proposed a social-historical model that captures the user's check-in history and short-term effect to explore user's check-in behavior. Lian et al. (2015) proposed an effective recommendation system by analyzing the historical check-in data.

Mining Human Mobility Data. Our work can be categorized into human mobility modeling. There are existing studies on human mobility modeling by exploiting mobility patterns to enable various applications (Zheng et al. 2014). The major stream of human mobility modeling is to perform latent human mobility clustering and to analyze the semantic meanings of mobility clusters, which has enabled us to discover various categories of outdoor activities (also named discovery of geographic topics and urban functions). The work along this stream are usually based on two types of data: (i) location based social networks (LBSNs) data (Kling and Pozdnoukhov 2012; Pozdnoukhov and Kaiser 2011; Yin et al. 2011) and (ii) mining GPS traces data (Qi et al. 2011; Fu et al. 2014; Yuan et al. 2015; Zheng 2015a; Long and Shen 2015; Zhao et al. 2015; Lee and Holme 2015). Giannotti et al. (2007) introduced the sequential pattern mining paradigm that analyzes the trajectories of moving objects and provided several different methods to acquire the patterns from trajectories data. Monreale et al. (2009) focused on predicting the next location of a moving object with a certain level accuracy. Zheng et al. (2010a) proposed a method based on supervised learning to automatically deduce users' transporation modes, such as walking and driving, from GPS logs. Ying et al. (2011) proposed a method to predict the next location of a user's movement based on both the geographic and semantic features of users' trajectories. Fu et al. (2014) exploited the

geographic dependencies of the value of an estate with a geographic method named ClusRanking. Shang et al. (2014) focused on predicting human mobility and detecting over-crowed stations in public transport networks by using human-tracking data, and then proposed a method based on network expansion to find unobstructed routes to go around these over-crowed stations. Wang et al. (2015) developed a hybrid model integrating both the regularity and conformity of human mobility to make a location prediction, which captured users' regular movement patterns and their occasional visits influenced by others. Lian et al. (2014) exploited the information of activity area vectors of users and influence area vectors of POIs to augment the modeling of users and POIs. Yin et al. (2014) constructed a temporal context-aware mixture model (TCAM) to explore the intentions and preferences by analyzing the user behaviors in social media systems. Fu et al. (2015) proposed a latent factor model to learn the profolio of community functions for real estate from human mobility data and conducted extensive experiments on real-world human mobility data.

Link Analysis and Topic Modeling. Topic modeling methods have received lots of attention in recent data mining research, which focused on Latent dirichlet allocation (LDA) (Blei et al. 2003). Blei and Lafferty (2006) proposed a correlated topic model (CTM), which can model topic correlation between topics. Blei and Lafferty (2007) developed a hierarchical topic model of documents that can capture the correlations between the topics and construct topic graph and applied the model to large document collections. Nallapati et al. (2008) presented two different models that can jointly model the text and citations by combining the ideas of LDA and Mixed Membership Block Stochastic Models. Liu et al. (2009) developed the Topic-Link LDA model that performs topic modeling and author community, which brings both topic modeling and community discovery in one unified model. The idea of topic models have also been adapted to model spatial-temporal data. Kling and Pozdnoukhov (2012) modeled urban dynamics using spatial-textual data from LBSNs data. Particularly, the authors developed a probabilistic topic model to learn a decomposition of location traces and obtain a set of urban topics related to citizen activities. Pozdnoukhov and Kaiser (2011) proposed a streaming Latent Dirichlet Allocation topic model for exploring spatial-temporal structures of the topical content in a stream available from LBSNs for sensing various aspects of evolution and dynamics of urban systems. Yin et al. (2011) proposed a Latent Geographical Topic Analysis (LGTA) that combines location based modeling and text based modeling for discovering different topics of interests that are coherent in geographical regions and for comparing topics across different geographical locations. Kim et al. (2015) developed a topic model to capture the semantic regions where people post messages with a coherent topic as well as the pattern of movement between the semantic regions. With GPS traces, more studies has been conducted based on human mobility patterns (e.g., direction, speed, time, location) for quantifying urban functions (Qi et al. 2011; Yuan et al. 2015; Zheng 2015a; Long and Shen 2015; Zhao et al. 2015; Lee and Holme 2015). For example, the work in Yuan et al. (2015) analogized human mobility patterns as words, and exploited both topic modeling and spectrum analysis to analyze the urban functions of regions.

7 CONCLUSIONS

With the development of GPS-enabled vehicles, large-scale vehicle traces such as taxicab GPS trajectories are increasingly available for individuals. Even though these data can inform where individuals are, the semantic meaning of the trip purpose is usually unknown. Indeed, identifying trip purposes can help understand people's mobility patterns, to better monitor the traffic flows between regions of different functionality in a city, and thus can guide urban planning. To that end, in this article, we have developed a probabilistic analysis framework for inferring the trip purposes of taxi passengers. The recorded O-D pairs from taxi GPS are augmented by the neighborhood POIs, and we further introduce a latent factor POI Topic to represent the mixed functionality of

a point in the city. Considering the POI Topics of the origins and destinations, along with the trip time, we proposed POI Link Model to generate the POI Links, which can reveal the semantic meanings of the trip purpose. In addition, we have experimented on the taxi data of New York City, supplemented by the POI data of the city. The results have shown that the model can uncover the trip purposes in terms of the POI Topics and the trip time, and we have also implemented quantitative experiments to predict destinations and trip time, which is also shown to outperform other baseline methods.

Future work can be extended in the following ways. First, O-D pairs can be augmented with more heterogeneous data such as latitude, longitude, as well as people's activities (e.g., check-in behaviors); Second, POI can be re-weighted differently according to the popularity and importance for a particular region, such that we can identify th e functionality and activity types more clearly. Furthermore, the model can be extended to capture the traffic flows between regions of various functionality.

REFERENCES

Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. 2008. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* 9, Sep (2008), 1981–2014.

David Blei and John Lafferty. 2006. Correlated topic models. *Adv. Neural Info. Process. Syst.* 18 (2006), 147.

David M. Blei and John D. Lafferty. 2007. A correlated topic model of science. *Ann. Appl. Stat.* (2007), 17–35.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, Jan. (2003), 993–1022.

Rinku Dewri, Prasad Annadata, Wisam Eltarjaman, and Ramakrishna Thurimella. 2013. Inferring trip destinations from driving habits data. In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*. ACM, 267–272.

Yanjie Fu, Yong Ge, Yu Zheng, Zijun Yao, Yanchi Liu, Hui Xiong, and Nicholas Jing Yuan. 2014. Sparse real estate ranking with online user reviews and offline moving behaviors. In *Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM'14)*. IEEE, 120–129.

Yanjie Fu, Guannan Liu, Spiros Papadimitriou, Hui Xiong, Yong Ge, Hengshu Zhu, and Chen Zhu. 2015. Real estate ranking via mixed land-use latent models. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 299–308.

Yanjie Fu, Hui Xiong, Yong Ge, Zijun Yao, Yu Zheng, and Zhi-Hua Zhou. 2014. Exploiting geographic dependencies for real estate appraisal: A mutual perspective of ranking and clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1047–1056.

Barbara Furletti, Paolo Cintia, Chiara Renso, and Laura Spinsanti. 2013. Inferring human activities from GPS tracks.

Huiji Gao, Jiliang Tang, and Huan Liu. 2012. Exploring social-historical ties on location-based social networks. In *Proceedings of the International Conference on Web and Social Media (ICWSM'12)*.

Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. 2007. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 330–339.

Younghoon Kim, Jiawei Han, and Cangzhou Yuan. 2015. TOPTRAC: Topical trajectory pattern mining. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 587–596.

Felix Kling and Alexei Pozdnoukhov. 2012. When a city tells a story: Urban topic analysis. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. ACM, 482–485.

Ravi Kumar, Mohammad Mahdian, Bo Pang, Andrew Tomkins, and Sergei Vassilvitskii. 2015. Driven by food: Modeling geographic choice. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. ACM, 213–222.

Minjin Lee and Petter Holme. 2015. Relating land use and human intra-city mobility. *PloS One* 10, 10 (2015), e0140152.

Defu Lian, Xing Xie, Vincent W. Zheng, Nicholas Jing Yuan, Fuzheng Zhang, and Enhong Chen. 2015. CEPR: A collaborative exploration and periodically returning model for location prediction. *ACM Trans. Intell. Syst. Technol. (TIST)* 6, 1 (2015), 8.

Defu Lian, Cong Zhao, Xing Xie, Guangzhong Sun, Enhong Chen, and Yong Rui. 2014. GeoMF: Joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 831–840.

Youfang Lin, Huaiyu Wan, Rui Jiang, Zhihao Wu, and Xuguang Jia. 2015. Inferring the travel purposes of passenger groups for better understanding of passengers. *IEEE Trans. Intell. Transport. Syst.* 16, 1 (2015), 235–243.

Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. 2009. Topic-link LDA: Joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 665–672.

Ying Long and Zhenjiang Shen. 2015. Discovering functional zones using bus smart card data and points of interest in Beijing. In *Geospatial Analysis to Support Urban Planning in Beijing*. Springer, 193–217.

Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. 2009. Wherenext: A location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 637–646.

Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. 2008. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 542–550.

Alexei Pozdnoukhov and Christian Kaiser. 2011. Space-time dynamics of topics in streaming text. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-based Social Networks*. ACM, 1–8.

Guande Qi, Xiaolong Li, Shijian Li, Gang Pan, Zonghui Wang, and Daqing Zhang. 2011. Measuring social functions of city regions from large-scale taxi behaviors. In *Proceedings of the 2011 IEEE International Conference on Pervasive Computing and Communications (PERCOM'11)*. IEEE, 384–388.

Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. 2012. Finding your friends and following them to where you are. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*. ACM, 723–732.

Shuo Shang, Danhuai Guo, Jiajun Liu, and Kuien Liu. 2014. Human mobility prediction and unobstructed route planning in public transport networks. In *Proceedings of the 2014 IEEE 15th International Conference on Mobile Data Management*, Vol. 2. IEEE, 43–48.

Xuerui Wang and Andrew Mccallum. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 424–433.

Yingzi Wang, Nicholas Jing Yuan, Defu Lian, Linli Xu, Xing Xie, Enhong Chen, and Yong Rui. 2015. Regularity and conformity: Location prediction using heterogeneous mobility data. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1275–1284.

Fei Wu and Zhenhui Li. 2016. Where did you go: Personalized annotation of mobility records. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 589–598.

Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. 2013. Semantic trajectories: Mobility data computation and annotation. *ACM Trans. Intell. Syst. Technol. (TIST)* 4, 3 (2013), 49.

Hongzhi Yin, Bin Cui, Ling Chen, Zhiting Hu, and Zi Huang. 2014. A temporal context-aware model for user behavior modeling in social media systems. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. ACM, 1543–1554.

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical topic discovery and comparison. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, 247–256.

Josh Jia-Ching Ying, Wang-Chien Lee, Tz-Chiao Weng, and Vincent S. Tseng. 2011. Semantic trajectory mining for location prediction. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 34–43.

Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. 2015. Discovering urban functional zones using latent activity trajectories. *IEEE Trans Knowl. Data Eng.* 27, 3 (2015), 712–725.

Kai Zhao, Mohan Prasath Chinnasamy, and Sasu Tarkoma. 2015. Automatic city region analysis for urban routing. In *Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW'15)*. IEEE, 1136–1142.

Vincent W. Zheng, Yu Zheng, Xing Xie, and Qiang Yang. 2010b. Collaborative location and activity recommendations with GPS history data. In *Proceedings of the International Conference on World Wide Web (WWW'10)*. 1029–1038.

Yu Zheng. 2015a. Methodologies for cross-domain data fusion: An overview. *IEEE Trans. Big Data* 1, 1 (2015), 16–34.

Yu Zheng. 2015b. Trajectory data mining: An overview. *ACM Trans. Intell. Syst. Technol. (TIST)* 6, 3 (2015), 29.

Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol. (TIST)* 5, 3 (2014), 38.

Yu Zheng, Yukun Chen, Quannan Li, Xing Xie, and Wei-Ying Ma. 2010a. Understanding transportation modes based on GPS data for web applications. *ACM Trans. Web (TWEB)* 4, 1 (2010), 1.

Zack Zhu, Ulf Blanke, and Gerhard Tröster. 2014. Inferring travel purpose from crowd-augmented human mobility data. In *Proceedings of the 1st International Conference on IoT in Urban Space*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 44–49.

Received November 2016; revised March 2017; accepted March 2017