Travel Purpose Inference with GPS Trajectories and Geo-tagged Social Media Data

Chuishi Meng*, Yu Cui†, Qing He†, Lu Su*, and Jing Gao*

*Department of Computer Science and Engineering, SUNY Buffalo
†Department of Civil, Structural and Environmental Engineering, SUNY Buffalo
Email:{chuishim, ycui4, qinghe, lusu, jing}@buffalo.edu

***Abstract*—In people's daily lives, travel takes up an important part and many trips are generated everyday, such as going to school or shopping. With the widely adoption of GPS-integrated devices, a large amount of trips can be recorded with GPS trajectories. These trajectories are represented by sequences of geo-coordinates and can help us answer simple questions such as "where did you go". However, there is another important question awaiting to be answered, that is "what did/will you do", i.e., the trip purposes inference. In practice, people's trip purposes are very important in understanding travel behaviors and estimating travel demands. Obviously, it is very challenging to infer trip purposes solely based on the trajectories, because the GPS devices are not accurate enough to pinpoint the venues visited. Fortunately, the functionality and POI popularity near trip end locations can give us hints on people's activities. In order to infer trip purposes, this paper proposes a dynamic Bayesian network model which incorporates three important factors: the sequential properties of trip activities, the functionality and POI popularity of trip end areas. In addition, the POI popularities are learned from nearby social media data, and we propose an efficient method with local candidate pools to identify POIs from geo-tagged social media messages. Extensive experiments are conducted on a trajectory data collected from 8,361 residents in the Bay area, and the POI popularity is mined from 6.9 million geo-tagged tweets. Experimental results demonstrate the advantages of the proposed method on correctly inferring the trip purposes.**

***Keywords*-Dynamic Bayesian Network, Social Media, Point of Interest, Trip Purpose**

I. Introduction

Nowadays, we are able to collect a large amount of human trajectories because of the ubiquitous adoption of GPS-integrated devices. For instance, smart phones can track real-time trajectories, geo-tagged messages on social media platforms, such as Twitter, can also reveal users' trajectories. On the surface, these human trajectories record people's daily trips with a sequence of geo-coordinates. But in essence, they reveal people's activities or trip purposes, e.g., "shopping" or "eat out". With these abundant trajectory data in hand, we are curious to ask "what did/will you do when you arrive at one place?". This is the trip purpose inference problem.

The inference of people's trip purposes has many benefits for the whole society. First, people's trip purposes can help government officials understand travel behaviors and estimate travel demands which will lead to better city planning and investment decisions. In addition, it can provide customers with more accurate recommendations and better services, and this recommendation can be made before users start their trips.

However, trip purpose inference is very challenging, because the GPS devices are not accurate enough to pinpoint the venues visited. Although there have been some attempts on this research topic, existing methods usually adopt traditional classification methods. These methods overlooked several important trip properties such that it makes them less practical. Firstly, people's activities usually follow certain patterns, and there are intrinsic relationships among the sequence of activities. Take the trips shown in Figure 1 as an example. Parents may drop off their children at school before going to work, and people would eat in a restaurant after shopping. Similar activity patterns commonly reside across different users. In such case, knowing other activities of the day can help us infer the past or future activities. Second, it has been revealed that the functionality of the trip end area is useful to help infer the trip purposes [1], and the points of interest (POI) are usually utilized to characterize the land use information. However, the functionalities of a location would not reveal how people like it. Obviously, not all the POIs attract equal attentions. In other words, some of them are more popular then the others. Because the trip purpose is a people-centric concept, the popularity of the venues would be useful for the inference. Unfortunately, there is no existing methods capture the POIs' popularities.

In this work, we propose to infer trip purposes with users' trajectories, POIs and social media messages near the trip end locations. In order to capture the sequential property of the trips, this paper proposes a Dynamic Bayesian Model in which the trip purposes

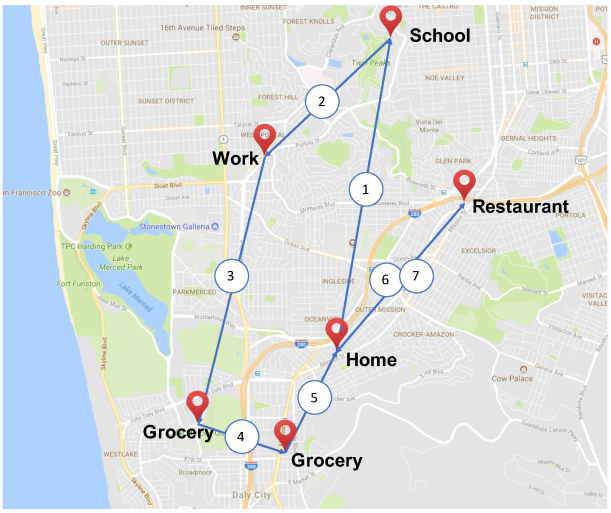


Figure 1: A typical user's daily trips

are hidden variables. By this means, we incorporate the knowledge of all other trips in the sequence to infer the trip purposes. Another advantage of adopting a Bayesian method is that we can derive ranked results of purpose inference with corresponding probabilities, e.g., 60% shopping, 35% eat out and 5% Recreation. The results with higher ranks are also helpful, especially for trips with vague purposes, such as having lunch while shopping in a mall. In addition, this paper proposes to mine the popularity of POIs from geo-tagged social media data. This information can provide a different perspective on the functionality of trip end locations, and it can help us infer the activities performed. However, the social media data is noisy and it is usually very hard to extract relevant information. In order to solve this problem, we propose an effective approach with local candidate pools to extract POI mentions from geo-tagged Twitter data. Then the popularity of a POI can be captured by the number of mentions in the social media.

In summary, this work makes the following contributions:

- We formulate the trip purpose inference problem with a Dynamic Bayesian Network which captures the sequential property of people's activities. As a result, a sequence of trips is considered as an integral part to infer a past or future trip purposes. In addition, the inferred ranked results can help us handle trips with vague purposes.
- We propose an effective method to extract mentioned POIs from geo-tagged social media messages, and model POIs' popularities in trip end locations. This popularity knowledge is useful with improving the inference performance.
- We conduct extensive experiments on real-world data with 8,631 trajectories and 6.9 million geo-tagged tweets in the Bay Area, CA. The results demonstrate the advantage of the proposed method on correctly inferring users' trip purposes.

The rest of the paper is organized as follows. We formally define the trip purpose inference problem in Section II. The proposed methods are detailed in Section III, and experiments are shown in Section IV. We review the related work in Section V then conclude the paper in Section VI.

II. Overview

In the following, we introduce several important concepts that will be used throughout the work, then formally define the trip purpose inference problem.

Definition 1. *A* Trajectory *(Tr) is a sequence of time-ordered spatial points, $Tr : l_1 \rightarrow l_2 \rightarrow \cdots \rightarrow l_n$ where each point l is represented by a pair of GPS coordinates, i.e., longitude and latitude.*

In this paper, we regard "trip" as the movement from one location to another, e.g., $l_1 \rightarrow l_2$, and we refer these GPS points l as "trip end locations". Take the trajectory in Figure 1 as an example. The trajectory comprises eight GPS points which follows the sequence of $l_{Home} \rightarrow l_{School} \rightarrow l_{Work} \rightarrow l_{Grocery} \rightarrow l_{Grocery} \rightarrow l_{Home} \rightarrow l_{Restaurant} \rightarrow l_{Home}$. From the perspective of trips, it has seven trips labelled by blue circles. However, in most cases, we cannot know the activities performed in a location without users' input. In other words, we don't know whether a user is shopping in a grocery store or having meal at a restaurant, given only the geo-coordinates of trip end locations.

Definition 2. Trip Purpose *is the activity that a user performed at a trip end location.*

As the name refers, it denotes the purpose of a trip. In the following sections, we use "trip purpose" and "activity" interchangeably. As shown in Table I, we categorize all trip purposes into eight categories, i.e., "Home", "Education", "Shopping", "EatOut", "Recreation", "Personal", "Work", and "Transportation". The example activities shown in Table I are defined by the California Household Travel Survey (CHTS) [2] which collected people's daily trajectories and activities. More details about this data set will be discussed in Section IV.

Definition 3. Point of Interest *(*POI*) is a specific location that someone may find useful. In this work, they represent venues in the physical world, e.g., banks and shopping malls. Each* POI *is associated with properties such as name, address, coordinates, category and etc.*

Table I: Activity Categories

Category	Example Activities
Home	Any activities performed at home
Education	School, Class, Laboratory, Meal at college, After-school sports activity, Library, Clubs, etc.
Shopping	Routine shopping (groceries, clothing, convenience store, etc.), Shopping for major purchases or specialty items (appliance, electronics, new vehicles, major household repairs, etc.), Service private vehicle (gas, oil, lubes, etc.)
EatOut	Drive through meals (snacks, coffee, etc.), Eat meal at restaurant/dinner
Recreation	Indoor or outdoor exercise, Sports, Health care
Personal	Household errands, Personal business (visit attorney, accountant, etc.), Civic/religious activities, Entertainment (movies, sporting events, etc.)
Work	All activities performed at the work place
Transportation	Change type of transportation (walk to bus, walk to parking lot, etc.), Pick up/drop off passengers

Definition 4. Geo-tagged Tweet *is a Twitter message associated with a pair of GPS coordinates where the message was generated.*

Problem Definition. **Given** the trajectories of users, the points of interest and the Twitter messages near trip end locations, our **Objective** is to infer the purposes of trips.

Note that some of the trips have labels with corresponding purposes. The labels can be provided by users through Apps, manually recorded by users' trip diary, or mined from Twitter messages. However, acquiring these labels is very difficult, and it is usually assumed to be unavailable for a large portion of trips.

III. Trip Purpose Inference with Dynamic Bayesian Network

A trip's purpose is determined by many factors, such as other activities of the day, the category of the visited venue, the functionality and the popularity of the destination area. Before discussing the proposed methods, we first shed some light on how these factors associate with the trip purpose inference.

Sequential activities of the day. Common sense tells us that users' activities usually follow some patterns, and there are intrinsic relationships among the sequence of activities. For instance, parents may drop off their children at school before going to work, people would eat in a restaurant after shopping in a mall, and etc. Similar patterns among sequential activities widely exist, and it is very useful information for the purpose inference.

The category of the visited venue usually correlates with the trip purpose. For example, people arriving at a restaurant are very likely to have lunch or dinner; checking in at a mall tells us he will be shopping. There will be close relations between the category of venue people visited and their trip purposes. Unfortunately, the GPS devices are not accurate enough to pinpoint the venues visited. In addition, it is also not easy to acquire this knowledge from people because of privacy concerns.

The functionality of the trip end area reveals the general usage of the nearby area. When we are not aware of the specific venue that a user visited, the nearby POIs can give us a hint about what the trip purpose would be. For example, arriving at a place with many shops nearby means people will go shopping with a higher probability. Specifically, the distribution of POI categories is a good feature to denote the functionalities of a location.

The popularity of the trip end area. Although the nearby POIs can help us understand the functionalities of a location, it cannot capture how people think about this area. Obviously, not all the venues attract equal attentions. In other words, some of them are more popular than the others. The popularity of the venues can be a useful feature for the purpose inference task, because the trip purpose is indeed a people-centric concept. Fortunately, social media can help us out here. Take Twitter as an example, people can send geo-tagged messages, and many of them contains the comments towards nearby POIs. By matching these geo-tagged tweets to real-world POIs, we can reveal venues' popularities accordingly.

A good trip purpose inference method needs to consider all these factors and the intrinsic relationships among them. In the following sections, we will first describe the proposed method for POI popularity modeling with social media data, then demonstrate the proposed Dynamic Bayesian Network approach.

A. POI Popularity Modeling

Based on the above reasoning, a POI's popularity can be captured if we can accurately identify them from tweet messages. The basic idea is that a POI is more popular if it has been mentioned by more tweets. However, this is a very challenging task. Firstly, social media data are very short. Existing named entity extraction methods perform poorly on these messages which have limited contexts. For example, "apple" may refer to the IT company or the fruit. Second, social media data are also very noisy. People usually use informal languages and names in tweets, such that we cannot expect to match a POI's full name in tweets. For instance, a tweet "dinner 2 tacos from lacorneta" mentions a restaurant "La Corneta Taqueria" without the full name, but rather with an abbreviation.

In this section, we propose a method to learn POIs' popularities from geo-tagged tweets. It can be much easier to identify mentioned POIs from these tweets because their associated geo-coordinates give us a good hint. Specifically, we can narrow down the search space to all nearby POIs. Compared with traditional methods [3] that works with a large POI knowledge base, our method can restrict the POI candidates to several dozens in a local area.

The proposed method works as follows. For each geo-tagged tweet, we first construct a local candidate pool with nearby POIs. Then a match index is calculated between the tweet content and each candidate POI names, and the POI with the highest index is marked as matched. Finally, the POI popularity of a trip end area can be derived by aggregating all the identified POIs from geo-tagged tweets. By this means, we can identify the mentioned POIs from geo-tagged tweets both effectively and efficiently. In the following, we describe each component in detail.

POI Local Candidate Pool Construction. In order to identify POI mentions for a geo-tagged tweet, we first construct a local candidate pool with all the POIs near the geo-coordinates of the tweet. Considering the accuracy of GPS devices, we shouldn't set the range to be too small or too large. In this work, we set the range as 200 meters which usually results in a pool with several dozen candidates.

Calculate POI Match Index. After constructed a local candidate pool for each tweet, the next step is to find the best matched POI among all candidates. To this end, we design a match index to measure the similarity between a tweet and a POI name. This match index considers two essential factors:

- The number of matched terms. The match index should be larger if there are more terms matched between a tweet and a POI name. For example, a tweet "Was just told by a teenager working at this Jamba Juice, that I looked like a young Walter White" mentions the POI "Jamba Juice Redwood City", and two terms are matched between them. However, there is another nearby POI "Geoff White Photographers" which matches a term "white" to the tweet. In this case, "Jamba Juice" with 2 matched terms should be weighed higher than the other one with only 1 matched term.
- The rareness of the matched terms. Some terms may frequently appear in the candidate pool. For example, there is no surprise that many POI names contain "San Francisco" in the Bay area. Then these terms should have less impact on the matching index. On the contrary, terms such as "Corneta" is relatively rare. In fact, this term only appears in a restaurant named "La Corneta Taqueria". No doubt that these terms should have larger impact on the matching index. In other words, if terms like "Corneta" matched between a tweet and a POI name, we should have a high belief that the tweet mentioned the restaurant "La Corneta Taqueria".

In this work, we propose a POI Match Index which characterizes the aforementioned factors. Specifically, each Tweet T_i is represented by a set of terms, i.e., $T_i = \{u_1, u_2, \cdots, u_m\}$. Similarly, each candidate POI's name is represented by $P_j = \{v_1, v_2, \cdots, v_n\}$. Then the set of Matched Terms MT between T_i and P_j are

$$MT(T_i, P_j) = T_i \cap P_j. \tag{1}$$

We can calculate the Match Index as follows

$$\begin{aligned} MI(T_i, P_j) &= |MT(T_i, P_j)| \\ &\times \log \frac{N_{pool}}{1 + \sum_{k=1}^{N} \mathbb{1}(MT(T_i, P_j) \in P_k)}, \end{aligned} \tag{2}$$

where N_{pool} denotes the size of the candidate pool. $\mathbb{1}(\cdot)$ is the indicator function which returns 1 if and only if the condition holds. In addition, $\sum_{k=1}^{N} \mathbb{1}(MT(T_i, P_j) \in P_k)$ calculates the frequency of the matched terms MT in the POI candidate pool. As shown in the equation 2, the first term considers the number of matched terms, and the second term considers the rareness of the matched terms. Note that $|MT(T_i, P_j)|$ could be zero, i.e., there are no matched terms between T_i and P_j. In this case, the Match Term Index will be 0 which is reasonable.

After calculated the Match Index between T_i and every P_j in its candidate pool, we can return the one with the highest index as the identified POI. However, we still need to set a threshold to the MI, and return non-identified if none of MIs exceed the threshold. In sum, With the constructed local POI candidate pool and the proposed Match Term Index, we can accurately identify the nearby POIs that mentioned in the Tweets.

POI Popularity Modeling. After extracted the POIs mentions from social media data, we can further represent the POI popularity across different categories by counting the corresponding mentions from social media. For example, if restaurants are mentioned by 10 different tweets, we will count 10 towards the popularity of the POI category "Food". Then the counts can be normalized into a distribution across all the POI categories in Table II.

B. Dynamic Bayesian Network Construction

In this work, we propose a Dynamic Bayesian Network (***DBN***) to model people's sequential activities. As shown in the Figure 2, $a \in A$ denotes the activity performed (or trip purpose), $c \in C$ denotes the category

Table II: POI Category

POI Category	Google Place Type
Money	accounting, atm, bank, post office, finance
Leisure	amusement park, aquarium, art gallery, casino, bowling alley, gym, movie rental, movie theater, museum, park, stadium, zoo
Food	bakery, cafe, food, meal takeaway, restaurant, meal delivery
Bar	bar, night club
Care	beauty salon, hair care, spa
Store	bicycle store, book store, clothing store, convenience store, department store, electronics store, florist, furniture store, grocery, supermarket, grocery or supermarket, hardware store, home goods store, jewelry store, liquor store, pet store, shoe store, shopping mall, store, car dealer
Trans	bus station, subway station, train station, taxi stand, parking, car rental, airport, light rail station, transit station
Auto	car repair, car wash, gas station
Religion	cemetery, church, funeral home, hindu temple, mosque, place of worship, synagogue
Civic	courthouse, lawyer, police, fire station, city hall, embassy, local government, local government office
Health	dentist, doctor, health, hospital, pharmacy, physiotherapist,veterinary care
Improve	electrician, locksmith, moving company, painter, plumber, real estate agency, travel agency, general contractor, roofing contractor, insurance agency, laundry, storage
Edu	library, school, university
Lodge	rv park, lodging, campground

of POI a user visited, and l is the trip end locations. All activities (A) and POI categories (C) defined in this paper are shown in Table I and Table II.

The ***DBN*** model can be interpreted in a generative process, or in other words, in a user's decision making process. For each trip i, a user first decides an activity a_i (or purpose) based on his previous one a_{i-1}. Then he chooses a venue category c_i based on this choice of activity. At last, he chooses a geo-location l_i to finally perform the activity a_i in venue c_i. This process continues until the last trip.

The likelihood function of the proposed ***DBN***model is as follows

$$P(a,c,l) = P(a_0)P(c_0|a_0)P(l_0|c_0) \cdot \left(\prod_{i=1}^{N} P(a_i|a_{i-1})P(c_i|a_i)P(l_i|c_i)\right), \quad (3)$$

where $P(a_i|a_{i-1})$ is the probability of the activity a_i given previous activity a_{i-1}, $P(c_i|a_i)$ is the probability of the visited POI category given current activity a_i,

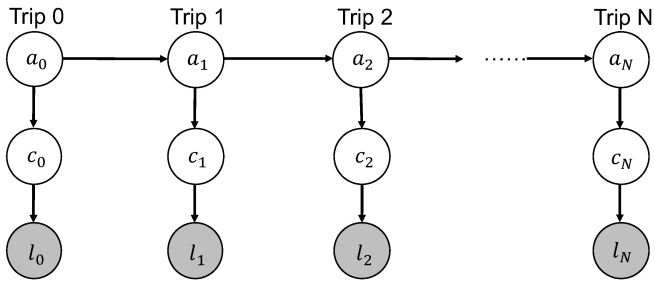


Figure 2: The Dynamic Bayesian Network

and $P(l_i|c_i)$ is the probability of chosen location l_i given currently chosen POI category c_i. In the proposed model, the visited location l_i is always observed, such that we can use Bayes's rule to approximate $P(l_i|c_i)$ as follows:

$$P(l_i|c_i) \propto \frac{P(c_i|l_i)}{P(c_i)} \propto \frac{P_{POI}(c_i|l_i)P_{tweet}(c_i|l_i)}{\sum_i P_{POI}(c_i|l_i)P_{POI}(c_i|l_i)} \times \frac{1}{P(c_i)}. \quad (4)$$

In the Equation 4, $P(c_i|l_i)$ denotes the POI category distribution given a geo-location l_i. This distribution is determined by two aforementioned factors: the functionality distribution $P_{POI}(c_i|l_i)$, and the popularity distribution $P_{tweet}(c_i|l_i)$. The first distribution is obtained from populating the nearby POIs, and the second distribution is obtained by extracting POI mentions from nearby tweets.

C. DBN Parameter Learning

There are two sets of parameters in the ***DBN***model: the transition probabilities $P(a_i|a_{i-1})$ and the emission probabilities $P(c_i|a_i)$. Note that in our problem, the activities a_i and visited venues c_i are not fully observed. In other words, many activities and corresponding venues are not labelled in the data. In order to learn the parameters from such incomplete data, we adopt the EM algorithm [4]. The process is summarized in Algorithm 1. It starts with an initial set of parameters. In each Expectation step (E-step), we compute the expected sufficient statistics for the parameter variables. Then in each Maximization step (M-step), we treat the expected sufficient statistics as observed, and perform Maximum Likelihood Estimation to estimate a new set of parameters. The algorithm continues between these two steps until converges.

D. DBN Prediction

With the learned parameters of the ***DBN*** mode, we can infer possible activities and their corresponding probabilities for any given trip. Specifically, we can calculate the posterior probability of the j_{th} activities given a user's trajectory $d \in \mathcal{D}$. This estimates the

Algorithm 1 DBN Parameter Learning

Input: DBN Structure $\mathcal{G}$, Partially observed trip data set $\mathcal{D}$, Initial set of Parameters $\theta_0 = \{P(a_i|a_{i-1}), P(c_i|a_i)\}$
Output: Learned DBN parameters θ_t

for $t \leftarrow 0, 1, \cdots$, until convergence **do**
  //Expectation-step
  for each $a \in A$ and each $c \in C$ **do** ▷ Initialize
    $M_t[a_i, a_{i-1}] = 0$
    $M_t[c_i, a_i] = 0$
  end for
  for each $d \in D$ **do**
    Run inference on the graph $\mathcal{G}$ using evidence d
    for each $a \in A$ and each $c \in C$ **do**
      $M_t[a_i, a_{i-1}] \leftarrow M[a_i, a_{i-1}] + P(a_i, a_{i-1}|d)$
      $M_t[c_i, a_i] \leftarrow M[c_i, a_i] + P(c_i, a_i|d)$
    end for
  end for
  //Maximization-step
  for each $a \in A$ and each $c \in C$ **do**
    $P_t(a_i|a_{i-1}) \leftarrow \frac{M_t[a_i, a_{i-1}]}{M_t[a_{i-1}]}$
    $P_t(c_i|a_i) \leftarrow \frac{M_t[c_i, a_i]}{M_t[a_i]}$
    $\theta_{t+1} \leftarrow \{P_t(a_i|a_{i-1}), P_t(c_i|a_i)\}$
  end for
end for
return θ_t.

probability of activity a_j out of all possible activities A, as shown in Equation 5.

$$P(a_j|d) = \frac{P(a_j, d)}{P(d)}, \forall a_j \in A, \tag{5}$$

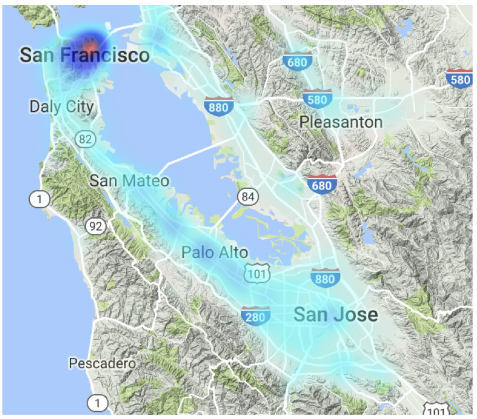
The returned results are possible activities ranked by their probabilities. Note that, generating a ranked result is a great advantage by adopting the Bayesian method, especially compared with traditional methods which can only provide a best guess. In fact, the top ranked inference results are very useful in real-world applications. Many classification tasks may have very vague decision boundaries, and usually the best guess results in poor performance. However, a ranked list with probabilities can help us identify several meaningful results and improve the inference accuracy. The experiments shown in Section IV-C provides a good demonstration.

IV. Experiments

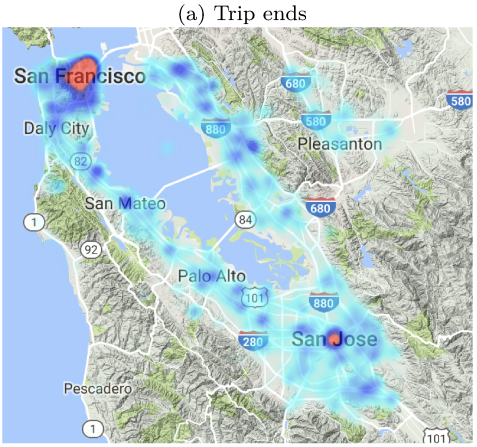
The proposed method is evaluated with real-world data sets including human trajectories, point of interests, and tweets in the Bay area, CA. In the following sections, we discuss the data sets, the baselines, and the evaluation results.

A. Data Sets

Trajectories. The California Household Travel Survey (CHTS) [2] collected travel information from residents across California's 58 counties. The survey was designed



(a) Trip ends



(b) Geo-tagged Tweets

Figure 3: Trip ends and Geo-tagged tweets

to obtain detailed information about the household socioeconomic characteristics and their travel behaviors. Among its various achievements, the survey collected 8,631 participants' GPS trajectories for a week. In addition, each trajectory is accompanied with a detailed trip diary which records visited POIs and trip purposes. The trip purposes are labelled by users with pre-defined categories, and some of them are shown on the second column in Table I. However, the diaries are far from exhaustive and many trips recorded by GPS devices are not logged in the diary. This is a common issue for traditional surveys because it requires too much efforts from the participants. The heat map of trip end locations are shown in Figure 3 (a).

Twitter Data. We collected 6.9 million geo-tagged tweets in the Bay area from Jan 31 2013 to Feb 16 2017. They are queried through Twitter APIs and filtered by geo-coordinates. The heat map of geo-tagged tweets are shown in Figure 3 (b).

Point of Interest. The POIs are queried through Google Places API [5]. In this work, we use its Nearby Search request to get the nearby POIs of any given geo-

Table III: Performance of Trip Purpose Inference

Accuracy	SVM	ANN	KNN	RF	DBN-top-1	DBN-top-2	DBN-top-3
Education	27.0%	41.9%	34.4%	40.2%	**52.3%**	**72.9%**	**78.4%**
Shopping	59.6%	65.3%	52.5%	78.6%	**80.1%**	**94.2%**	**98.4%**
EatOut	4.8%	30.0%	32.1%	60.6%	**79.0%**	**83.2%**	**85.5%**
Recreation	18.6%	39.0%	30.6%	55.5%	**62.7%**	**77.4%**	**84.7%**
Personal	7.7%	21.7%	22.5%	**50.4%**	42.2%	**65.0%**	**90.0%**
Transportation	**84.6%**	74.2%	59.1%	75.4%	68.3%	**84.3%**	**89.6%**
Average	33.7%	45.4%	38.5%	60.1%	**64.1%**	**79.5%**	**87.8%**
F1 Score	SVM	ANN	KNN	RF	DBN-top-1	DBN-top-2	DBN-top-3
Education	0.362	0.461	0.358	0.419	**0.484**	**0.757**	**0.840**
Shopping	0.517	0.600	0.452	0.756	**0.754**	**0.853**	**0.909**
EatOut	0.087	0.349	0.320	0.635	**0.712**	**0.835**	**0.879**
Recreation	0.281	0.430	0.342	0.585	**0.592**	**0.721**	**0.826**
Personal	0.134	0.299	0.266	**0.550**	0.476	**0.737**	**0.920**
Transportation	0.625	0.655	0.588	0.722	**0.735**	**0.869**	**0.919**
Average	0.334	0.465	0.387	0.611	**0.626**	**0.795**	**0.882**

coordinates, and the returned POI names and types are utilized by the proposed method. Some Google Place types are shown in the Table II.

B. Baselines

In the following experiment, we compare the proposed method with state-of-the-art methods for the trip purpose inference.

- Random Forest (***RF***) is an ensemble learning method for classification. It is constructed by a large amount of decision trees with sub-samples. After training, the prediction result for one record is decided by taking the majority vote across all decision trees.
- Support Vector Machine (***SVM***) tries to find the optimal decision boundary with the largest margins to classify data from different classes.
- Artificial Neural Network (***ANN***) can be trained to perform classification tasks. In the experiment, we adopt Multi-layer Perceptron, one typical kind of ANN, to predict the trip purpose.
- K-nearest Neighbor (***KNN***) finds a predefined number of training samples closest in distance to the new point, and predict the label from these samples.

In the following, we perform extensive experiments on the proposed ***DBN*** method, and demonstrate its advantages compared with baselines. Specifically, the performance of trip purpose inference is discussed in Section IV-C, the proposed POI mention identification method with social media is evaluated in Section IV-D, and we discuss how POI popularity features can affect the performance of trip purpose inference in Section IV-E.

C. Trip Purpose Inference

To compare the performance of all the methods on the trip purpose inference, we conduct experiments on the collected real-world data set. The features used for training include travel mode, previous activity category, activity time and duration, nearby POI category distribution, and nearby POI popularity distribution. We randomly select 80% trips as training data, and leave the rest for testing. All the methods are evaluated by 10 times, and the average results are reported. In addition, there are about one third trip end locations in the dataset are either home or work, because the survey collected trips from people's daily lives. Using these trips to train the model would greatly bias the performance. As a result, we only test the inference results on other activities, and assume the home and work trips are known. In other words, we are interested in non-trivial tasks of inferring non-home and non-work trip purposes.

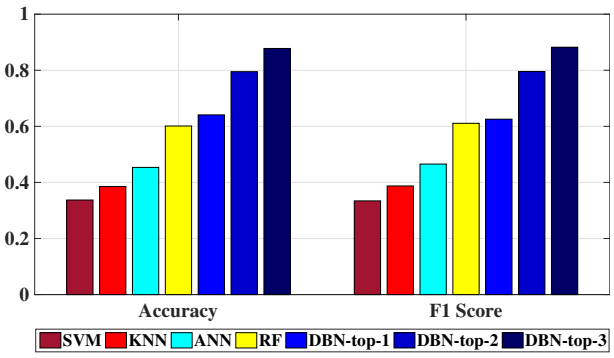


Figure 4: Average Performance of Trip Purpose Inference

Note that the proposed ***DBN*** model can output a ranked list of activities with probabilities. This is a great advantage by adopting the Bayesian method. As we have discussed in Section III-D the top ranked inference results are very useful in real-world applications, especially when the classification boundaries among classes

Table IV: Performance of POI Extraction from geo-tagged Tweets

Example Locations	Total Nearby Tweets	POI-mentioned Tweets	App-formatted Tweets	POI-identified Tweets
37.4028125, -121.881519784	575	184	116 (63.0%)	159 (86.4%)
37.3896875, -122.03080036	294	103	80 (77.7%)	93 (90.3%)
37.3278125, -121.8842176	1188	368	247 (67.1%)	305 (82.9%)
37.6715625, -122.472796763	1232	596	446 (74.8%)	495 (83.1%)
Average Recall	-	-	68.8%	83.5%

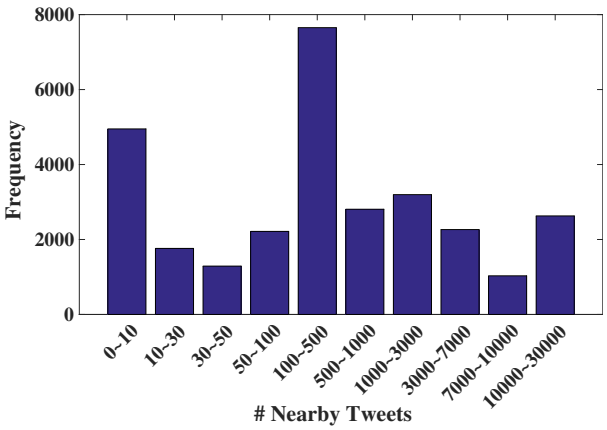


Figure 5: Histogram of tweets near trip ends

are vague. As a result, we also evaluate the ***DBN*** performance with top-2 and top-3 inference results in the experiment, denoted as DBN-top-2, and DBN-top-3 respectively. In these cases, we regard the inference as correct if the ground truth activity is among the top-2 or top-3 results.

The inference accuracy and F1 scores are shown in Table III, and the average results are also compared in Figure 4. We can observe that the proposed ***DBN*** model, including DBN-top-1, DBN-top-2 and DBN-top-3, outperforms other baselines on almost every activity category by higher accuracy and F1 score. This is because ***DBN*** model captures the intrinsic relationships among sequential activities, trip end locations' POI distributions and the popularities identified from the Twitter data. On average, the ***DBN*** model can reach 64% accuracy with the top-1 inference result. Moreover, the accuracy of top-2 and top-3 ranked results can reach 79.5% and 87.8%. These results are impressive and they demonstrate that the top-ranked results generated by the ***DBN*** model is very useful in the trip purpose inference.

D. POI Mention Extraction from Geo-tagged Tweets

In Section III-A, we propose to extract mentioned POIs from geo-tagged tweets. For each geo-tagged tweet, we can accurately identify the mentioned POI with a local candidate pool and the match index.

In practice, it is very hard to evaluate the performance of POI mention extraction from tweets, because we have so many nearby tweets and trips in the data set. Figure 5 shows the histogram of tweets near trip ends. On average, there are 2607 geo-tagged tweets near each trip end location (within 200 meters). Actually, it is impossible to be evaluated without a standard data set labelled by human workers. To this end, we recruited volunteers to label the POI mentioned tweets near 50 random trip end locations. Given a list of tweets, the volunteers are asked to judge whether those tweets mentioned any nearby POI, and whether the identified POIs are correct. In Table IV, we present the results on several example trip end locations and the average performance. For each location, we populate the total number of nearby tweets, the number of POI-mentioned tweets which are identified by human workers. Some of the tweets are formatted by third party Apps, such as Foursquare and Instagram. We can easily parse POIs from these well formatted tweets, for example, "Im at Applewood Pizza in San Carlos Ca" and "Bagel time! @ Bagel Street Cafe Town Center Alameda". However, there are still many POI-mentioned tweets without these formats. For instance, "Catch us at amc Mercado 20 12 am insurgent!!!!". The better performance shown in Table IV indicates the proposed method can also extract mentioned POIs from free tweets.

E. Impact of the POI Popularity Features

For each trip end, we can model its nearby POI's popularity with the social media data in Section III-A. In this section, we perform experiments to evaluate how this popularity distribution can help us infer people's trip purposes. We first evaluate its impact on the Random Forest model. As observed in Table III, the Random Forest model can also provide meaningful purpose inference, and it even has better performance on inferring the "Personal" activities. In this experiment, we train two Random Forest models, one with the POI popularity features (RF-Tweet) and the other without (Rf-noTweet). As shown in Table V. the POI popularity features mined from the Twitter data can improve the inference accuracy on every class.

In addition, we further evaluate the POI popularity features on the ***DBN*** model. Two ***DBN*** models

Table V: Accuracy Comparison with POI Popularity Features

Accuracy	RF-noTweet	RF-Tweet	DBN-noTweet	DBN-Tweet
Education	40.3%	**40.5%**	50.0%	**52.3%**
Shopping	77.0%	**78.6%**	**82.9%**	80.1%
EatOut	60.0%	**60.6%**	78.9%	**79.0%**
Recreation	53.2%	**55.5%**	**65.7%**	62.7%
Personal	47.7%	**50.4%**	38.9%	**42.2%**
Transportation	75.1%	**75.4%**	**69.5%**	68.3%

are trained and compared. One without the popularity features (DBN-noTweet), and the other with the popularity features (DBN-Tweet). The results are shown in the Table V. Interestingly, this popularity feature has different impacts on different activity categories. Specifically, incorporating the popularity features leads to higher accuracy on the inference of "Education", "EatOut", and "Personal". But it results in slightly lower accuracy on the inference of "Shopping", "Recreation", and "Transportation". The reason lies in the difference properties between the cyber and physical worlds. Take the activity "Education" as an example. One educational institute, e.g., high school, would be surrounded by many other POIs, e.g., restaurants. In this case, the weight of high school will be under-estimated in the POI distribution. Fortunately, we can identify more people's tweets in the school than in the restaurant, i.e., the school is more popular than other POIs in the social media. By this means, the POI's popularity feature can help us recognize the importance of the school. On the other hand, the POI's popularity may fail to capture the importance of certain POIs if there are not enough people discussing them in the social media, and this may result in slightly worse inference performance, such as in "Recreation" and "Transportation"'.

V. Related Work

There are several research fields related to this work, and we summarize them in this section.

Trajectory Mining. The ubiquitous adoption of GPS-integrated devices has enabled extensive studies on human trajectory data mining [6], [7], [8], [9]. Wu et al. [10] proposed a Markov Random Field model to infer the visited POIs given users' trajectories. It aims to answer the question "if a person is observed at certain location and time, which venue is the true destination of this person". In [9], Zhang et al. adopt Hidden Markov model to capture group-level human mobilities with social media data, and Latent activity states are represented by topical words. In order to capture the topical information of trajectories, Kim et al. [11] proposed a probabilistic model to cluster different trajectories. With their method, significant movement patterns that appear frequently in data can be recognized. Wu et al. [12] used Kernel Density Estimation to annotate trajectories with relevant terms from social media data. In addition, there have been studies [13], [14], [15] to extract the patterns underlying people's activities. Our task is different from these studies, as none of them attempts to capture the high-level trip purpose with both the knowledge from POI and social media.

The problem of trip purpose prediction [16], [1] has been mostly studied in the transportation field. They usually adopt traditional classification methods, such as rule-based methods [16], Random forest [1], Artificial neural network [17], decision tree [18] and etc. In comparison, this work proposes a novel Dynamic Bayesian model to infer the trip purposes.

POI Identification from Social Media Data. As we have mentioned, social media messages are quite noisy. This makes it very difficult to perform Named Entity Recognition from such short and noisy texts. In order to extract fine-grained location information from tweets, Li et al. [19] proposed a Conditional Random Field model to identify POI mentions from social media messages, and they further proposed a method [20] to link the POI name with Foursquare inventory. In [21], Flatow et al. proposed a data-driven approach to identify phrases associated with regions. Then it is used to label non-geotagged tweets with a regional area. In addition, a supervised Bayesian Model [22] is proposed to annotate POIs with tweet information. The above research aim at identify POIs from general social media messages, especially for tweets without geo-tags. However, in this work, we propose to extract POI mentions from geo-tagged tweets. The proposed method with local POI candidate pools can solve this problem effectively and efficiently.

VI. Conclusions

In this paper, we propose to infer people's trip purposes with their trajectory data. The knowledge of people's daily activities is very useful which can benefit both the government and residents. In order to accurately infer people's trip purposes, we propose a dynamic Bayesian network model for the sequential trips. This model can capture the intrinsic relationships among activities. In addition, we also propose to incorporate POIs' popularity information near trip end locations. This information can give us good hints about people's activities. Moreover, in order to deal with the noisy social media data, we propose an efficient method with local POI candidate pool to identify POIs from geo-tagged tweets. Extensive experiments were conducted on real-world data sets, and the results demonstrate advantages of the proposed method on accurately inferring the trip purposes.

References

[1] A. Ermagun, Y. Fan, J. Wolfson, G. Adomavicius, and K. Das, "Real-time trip purpose prediction using online location-based search and discovery services," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 96–112, 2017.

[2] "California household travel survey." [Online]. Available: http://www.dot.ca.gov/hq/tpp/offices/omsp/statewide_travel_analysis/chts.html

[3] W. Hua, K. Zheng, and X. Zhou, "Microblog entity linking with social temporal context," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, 2015, pp. 1761–1775.

[4] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[5] "Google places api." [Online]. Available: https://developers.google.com/places

[6] Y. Zheng, "Trajectory data mining: an overview," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, p. 29, 2015.

[7] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1082–1090.

[8] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in foursquare." *ICwSM*, vol. 11, pp. 70–573, 2011.

[9] C. Zhang, K. Zhang, Q. Yuan, L. Zhang, T. Hanratty, and J. Han, "Gmove: Group-level mobility modeling using geo-tagged social media," in *KDD: proceedings. International Conference on Knowledge Discovery & Data Mining*, vol. 2016. NIH Public Access, 2016, p. 1305.

[10] F. Wu and Z. Li, "Where did you go: Personalized annotation of mobility records," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 589–598.

[11] Y. Kim, J. Han, and C. Yuan, "Toptrac: Topical trajectory pattern mining," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 587–596.

[12] F. Wu, Z. Li, W.-C. Lee, H. Wang, and Z. Huang, "Semantic annotation of mobility data using social media," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 1253–1263.

[13] D. Hristova, M. J. Williams, M. Musolesi, P. Panzarasa, and C. Mascolo, "Measuring urban social diversity using interconnected geo-social networks," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 21–30.

[14] Q. Yuan, W. Zhang, C. Zhang, X. Geng, G. Cong, and J. Han, "Pred: Periodic region detection for mobility modeling of social media users," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 2017, pp. 263–272.

[15] C. Zhang, J. Han, L. Shou, J. Lu, and T. La Porta, "Splitter: Mining fine-grained sequential patterns in semantic trajectories," *Proceedings of the VLDB Endowment*, vol. 7, no. 9, pp. 769–780, 2014.

[16] L. Shen and P. R. Stopher, "A process for trip purpose imputation from global positioning system data," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 261–267, 2013.

[17] G. Xiao, Z. Juan, and C. Zhang, "Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 447–463, 2016.

[18] M. Oliveira, P. Vovsha, J. Wolf, and M. Mitchell, "Evaluation of two methods for identifying trip purpose in gps-based household travel surveys," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2405, pp. 33–41, 2014.

[19] C. Li and A. Sun, "Fine-grained location extraction from tweets with temporal awareness," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 43–52.

[20] Z. Ji, A. Sun, G. Cong, and J. Han, "Joint recognition and linking of fine-grained locations from tweets," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 1271–1281.

[21] D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich, and Y. Kanza, "On the accuracy of hyper-local geotagging of social media content," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 2015, pp. 127–136.

[22] K. Zhao, G. Cong, and A. Sun, "Annotating points of interest with geo-tagged tweets," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 417–426.