Transportation Research Part C 44 (2014) 363–381

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Urban activity pattern classification using topic models from online geo-location data

Samiul Hasan, Satish V. Ukkusuri *

Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

Article history:
Received 8 January 2014
Received in revised form 8 April 2014
Accepted 9 April 2014

Keywords:
Activity pattern classification
Activity-based modeling
Social computing
Location-based data
Big data
Social media
Topic modeling
Machine learning

ABSTRACT

Location-based check-in services in various social media applications have enabled individuals to share their activity-related choices providing a new source of human activity data. Although geo-location data has the potential to infer multi-day patterns of individual activities, appropriate methodological approaches are needed. This paper presents a technique to analyze large-scale geo-location data from social media to infer individual activity patterns. A data-driven modeling approach, based on topic modeling, is proposed to classify patterns in individual activity choices. The model provides an activity generation mechanism which when combined with the data from traditional surveys is potentially a useful component of an activity-travel simulator. Using the model, aggregate patterns of users' weekly activities are extracted from the data. The model is extended to also find user-specific activity patterns. We extend the model to account for missing activities (a major limitation of social media data) and demonstrate how information from activity-based diaries can be complemented with longitudinal geo-location information. This work provides foundational tools that can be used when geo-location data is available to predict disaggregate activity patterns.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

The introduction of location-based services in smartphone applications has enabled people to share their activity related choices (typically via "check-in") in their online social networks (e.g. Facebook, Foursquare, Twitter, etc.). Market analysis predicts 1.75 billion smartphones in the world by 2014 (eMarketer, 2014). These smartphones are typically equipped with ubiquitous location-based technologies suitable for check-in services offered by most of the social media applications. A survey on smartphone usage in the USA found that nearly one in five smartphone users are tapping into check-in services like Facebook Places, Foursquare (Comscore, 2011). The research community is realizing the potential to harness the rich information provided by these ubiquitous devices. The availability of this data has the potential to impact many areas including mobility and activity behavior analysis (Cheng et al., 2011; Noulas et al., 2011), marketing (Gao et al., 2012; Collins et al., 2013), social network analysis (Cho et al., 2011), urban planning (Cranshaw et al., 2012), and health monitoring (De Choudhury et al., 2013). As such, there are opportunities to develop fundamental tools to analyze this data at various levels of spatial and temporal resolution. Transportation researchers, also, have realized the potential of this data for travel demand modeling and analysis (Collins et al., 2013; Hasan, 2013; Cebelak, 2013; Hasan et al., 2013; Jin et al., 2014; Yang

* Corresponding author. Tel.: +1 7654942296.
E-mail addresses: samiul.hasan@gmail.com (S. Hasan), sukkusur@ecn.purdue.edu (S.V. Ukkusuri).

http://dx.doi.org/10.1016/j.trc.2014.04.003
0968-090X/© 2014 Elsevier Ltd. All rights reserved.

et al., 2014; Ni et al., 2014; Alesiani et al., 2014). A key challenge however is the lack of appropriate methodologies to handle such large data including the ability to address the limitations of this data.

The focus of this work is to *develop methodologies to understand individual activity patterns using large-scale location based data* obtained from social media check-in services. Through these services, individuals share their activities with the specific information on geo-location and timing of where and when they participate in those activities. Also, geo-location data can be collected in large-scale by recording the GPS coordinates from smartphones. Activity types and time of participation over multiple days can be observed from geo-location information either shared in online social media or collected by smartphones. This large-scale geo-location data can be useful to understand human activity behavior due to the extensive coverage that was unimaginable before. With the availability of new data sources like social media check-in services and smartphone GPS devices, there is a profound interest in understanding and modeling individual activity behavior. As such, geo-location data from these emerging sources has created opportunities to develop complex probabilistic models inferring the patterns of activity behavior. This paper investigates the idea of using large-scale geo-location data to infer individual activity patterns. To the best of our knowledge, this is one of the first methodological works for classifying user activity patterns using large-scale social media data.

Geo-location data from online social media and GPS devices has the potential to understand activity behavior in urban areas. Until now, many of the studies in activity-travel analysis have relied primarily on small-scale but detailed records of individual activity participation and have correlated socio-demographic information with activity participation behavior. Such modeling approaches have enormous importance in long-term policy-level analysis and planning applications. However, with the availability of big data such as the data from social media and smartphones we can observe activity participation of a large number of people over many days. We can analyze these observations to infer activity patterns up to the level of an individual without drawing any correlation with the socio-demographic attributes. Our analysis is motivated by three major shifts in our thinking about traditional approaches of activity analysis. *First*, geo-location data from a large number of individuals over many days provides the ability to analyze vast amounts of data instead of settling for small-scale data sets – accepting "the unreasonable effectiveness of data" (Halevy et al., 2009). *Second*, this approach embraces the real-world messiness in the data (e.g., missing observations) rather than depending on comprehensive data and develops methods that can account for the noisiness in these data sets. *Third*, it focuses on predicting behavior rather than explaining behavior or drawing correlations between individual activity participation and socio-demographic attributes.

Such an approach is very useful, particularly for geo-location data, to provide collective and individual patterns. This analysis may help to measure the travel demand of a region on a short-term basis; for instance, using these patterns we can compute the real-time origin–destination matrices for transportation operation models. In the context of travel demand analysis, large-scale geo-location data can provide valuable information in addition to the data from traditional surveys, and, as such, it does not replace the existing data sources but strongly complements them. Using correct analytics, such large-scale geo-location data can be used to gather complementary insights and will lead to a transformative understanding of urban travel behavior.

Geo-location data comes with larger sample size for longer period (e.g., for a year) without any significant costs and provides the location and timing of individual activity participation. However, three key limitations limit the use of traditional econometric tools for these data sets. These limitations include: (i) it does not have the detail descriptions of the activities as the start times and the end times of the activities are not reported; thus most of the current methodological approaches for modeling individual time-use behavior are not appropriate for this data; (ii) individuals are recognized by only the identification numbers without any detailed information on individual socio-economic characteristics (e.g., income, age, race, etc.); (iii) the data has missing activities, since we observe only the activities that an individual shares in social media.

However, recent advancements in machine learning techniques have made it possible to analyze large-scale geo-location data to find the spatio-temporal patterns. Specifically, hierarchical mixture modeling (popularly known as topic modeling) has emerged as a powerful methodological approach to find patterns and structure in large collections of data. These models find the latent patterns from a collection of data points where a pattern means a probability distribution over a set of predefined items. In the beginning, these topic models were used to find the underlying patterns or topics of words from a large-collection of documents (Blei et al., 2003). These patterns can be used for clustering, searching, summarizing and predicting a large corpus of documents. Later topic models were used to find patterns in images, audio and speech, genetic data, computer code and mobile phone location-sequence data. In this paper, we present an activity pattern recognition model based on a topic modeling approach.

Specifically, we find the embedded patterns found in a collection of individual activities. Such an approach can be used to classify urban activity patterns- a topic of interest in activity-travel analysis for a long time (Pas et al., 1982; Recker et al., 1985; Joh et al., 2001, 2002; Wilson, 2008; Allahviranloo and Recker, 2013). Although predicting activity travel behavior is one of the major objectives of activity-based modeling, activity pattern recognition or classification provides the basis for more theoretical or empirical analysis (Joh et al., 2001). Most of the previous approaches (Pas et al., 1982, 1983; Koppelman and Pas, 1983; Recker et al., 1985; Joh et al., 2002), used activity patterns as predefined and mainly focused on similarity measures between patterns for two reasons. *First*, typically in activity-travel behavior analysis individual activities are correlated with the socio-demographic and/or spatial contexts (Hanson and Hanson, 1981; Pas, 1984; Allahviranloo and Recker, 2013). Therefore similarities among the observed activity patterns are used to classify individuals so that representative activity patterns can be correlated with individual characteristics. *Second*, similarity values can be used as goodness-of-fit statistics to measure how well an activity-based model can predict the observed activity patterns

(Joh et al., 2002; Sammour et al., 2012). With the availability of passively generated activity information from data sources such as smartphone GPS devices and social media there is a renewed interest in activity pattern recognition methods. In addition to serving the above purposes, such methods can measure the differences between traditional survey data and geo-location data.

The proposed method for activity pattern classification/recognition has several differences from the previous approaches. *First*, the previous approaches used predefined activity patterns limiting the opportunity of the data to inform emerging patterns. The method adopted in this paper, however, avoids using predefined patterns and lets the data inform the patterns. *Second*, most of the previous activity classification methods require a complete sequence of activities participated by an individual. Since a complete sequence of activities are not available from geo-location data due to the missing activities, these approaches are not applicable to these data sets. *Third*, the proposed approach provides a generating mechanism to derive the patterns whereas such mechanisms were not available for previous approaches. A major advantage of a generative mechanism is that it can be extended to simulate activities based on input data. *Fourth*, the proposed method attaches a probability value with each item in a pattern giving a probabilistic explanation to the derived patterns.

Due to the ubiquity of smartphones and location-based services in social media, it is likely that more geo-location data will be available in near future. Activity-based travel behavior analysis can take advantage of the opportunities offered by such big data sources. The applications of the proposed modeling approach will be important due to the inherent limitations of geo-location data from social media. This paper is a first step towards developing such methodologies with the following specific contributions:

1. We demonstrate the use of a large-scale geo-location data set to analyze and understand individual activity patterns. It uses geo-location data from social media to infer and classify activity patterns. While such kind of geo-location data has been used in the field of ubiquitous computing and social media research, we add a new dimension to this type of data by inferring activity types based on venue categories.
2. We identify the major limitations of geo-location data for modeling activity behavior and propose methodological approaches to account for such limitations.
3. We adopt a novel methodological approach, inspired from machine learning research, appropriate for geo-location data to discover general and user-specific activity patterns. We present a topic model suitable to extract useful information on activity patterns without the socio-demographic attributes of the individuals.
4. We further adopt an extension to the proposed topic model to account for the missing activities so that a complete activity sequence can be generated.

2. Data collection

2.1. Data set

The data set used in this analysis is collected from Twitter, a widely used social media application, where users can post short messages up to 140 characters. These short messages are generally called status message in social media norms and specifically called "Tweets" in Twitter. When permitted by the users, their tweets are attached with geo-location information. In addition to posting status messages, Twitter allows its users to post statuses from third-party check-in services (e.g., Foursquare). When Foursquare users check-in to a place this status can be posted to their Twitter pages. In this work, we use a large-scale check-in data available from Cheng et al. (2011). The data set contains check-ins from February, 25, 2010 to January, 20, 2011.

After collecting the tweets we preprocessed the data where each data point is stored as a tuple with the following information:

tweet (tweetID)= {userID, screen name, tweetID, date, location, text}

An example of a tweet with a check-in looks like:

tweet (79132591248261120)={189872633, ####, 79132591248261120, Fri Jun 10 10:27:34 + 0000 2011, 40.7529422, −73.9780177,

"I'm at Central Cafe & Deli (16 Vanderbilt Ave., New York) http://4sq.com/jMS87x"}

From the original check-in data set, we select a subset of all the observations within New York City by creating a boundary region and retain all the check-in observations within that region. This step results in a data set of 20,606 users from New York City. Descriptive statistics on the number of check-in activities include: average = 33.03; min = 1; max = 1010; and standard deviation = 77.82. Fig. 1 shows the cumulative distribution of the check-in activities. However, to classify individual activity patterns we study only the users with at least 50 check-ins. Basic information about the data set is given in Table 1.

2.2. Identification of activity categories

One of the major advantages of using social media check-in data is the ability to identify activity purposes. Each check-in observation reports a short link to the original check-in source (e.g., Foursquare). When queried to the source, this link gives

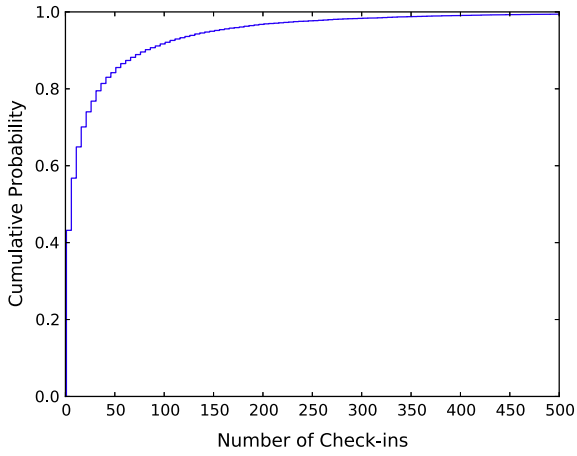


Fig. 1. Cumulative distribution of the number of check-in activities.

Table 1
Data set details.

Original data set	
Number of users	20,606
Number of check-ins	680,564
Study sample	
Number of geo-active users	3256
Number of check-ins by the geo-active users	504,000

information about the category of the visited location. We classify different activity categories based on the type of the visited locations (see Table 2). About 94.5% of the check-ins had a location category information available; for the rest, the check-in application could not resolve the corresponding location categories.

Table 2
Activity category classification.

Activity category	Type of visited location
Home (Ho)	Home (private), Residential Building (Apartment/Condo)
Work (Wo)	Office, Coworking Space, Tech Startup, Design Studio
Eating (Ea)	Coffee Shop, Restaurant, Pizza, Burger, Bodega, Caf, Diner, Sandwich, Bakery, Joint, Breakfast, Food, Bagel Shop, Steakhouse, Dessert Shop, Donut Shop, Ramen or Noodle House, Cupcake Shop, Gastropub, Taco Place, Gourmet Shop, Tea Room, Salad Place, Soup Place, College Cafeteria, Snack Place, Cafe, Winery, Fish and Chips Shop
Entertainment (En)	Pub, Entertainment, Venue, Nightclub, Bar, Theater, Club, Event Space, Stadium, Concert Hall, Speakeasy, Arcade, Other Nightlife, Dance Studio, Opera House, Casino
Recreation (Re)	Park, Gym, Other Great Outdoors, Playground, Bowling Alley, Dog Run, Scenic Lookout, Yoga Studio, Beach, Spa or Massage, Lake, Monument or Landmark, Historic Site, Other Great Outdoors, Zoo or Aquarium, Garden, Golf Course, Track, Field, Pool, Hiking Trail, Martial Arts Dojo, Basketball Court, Skating Rink, Surf Spot, Tanning Salon, Racetrack, Rest Area, Tennis Court, Resort, Hot Spring, Ski Area, Soccer Field, Campground, Great Outdoor, River, Hockey Arena, Vineyard, Ski
Shopping (Sh)	Supermarket, Store, Plaza, Pharmacy, Bookstore, Cosmetics Shop, Mall, Farmers Market, Boutique, Miscellaneous Shop, Automotive Shop, Food & Drink Shop, Flea Market, Sporting Goods Shop, Wine Shop, Bike Shop, Gift Shop, Record Shop, Hobby Shop, Butcher, Smoke Shop, Bookstore, Cheese Shop, Antique shop, Bridal Shop, Board Shop, Flower Shop, Optical Shop, Fish Market, Mobile Phone Shop
Social Service (So)	Hotel (not Hotel Bar), Building (not College Academic Building or College Science Building), Multiplex, Barbershop, Bank, Church, Hospital, Doctor's Office, Medical Center, Post Office, Laundromat or Dry Cleaner, Convention Center, Brewery, Gas Station or Garage, Conference Room, Courthouse, Tattoo Parlor, Voting Booth, Dentist, City Hall, Synagogue, Meeting Room, Roof Deck, Parking, hostel, Police Station, Emergency Room, Fraternity House, Religious Center, Embassy or Consulate, Temple, Cemetery, neighborhood, Veterinarian, Fire Station, Mosque, Motel, Farm, lighthouse, Auditorium, Sorority House, Subdivision or Housing Development, Nail Salon, Car Dealership, Animal Shelter, Distillery, Factory, Shrine, Lounge
Education (Ed)	University, College Academic Building, High School, Museum, Library, College (not College Gym, College Cafeteria, College Bookstore), School, Student Center, Planetarium

Identifying activity categories based on the types of check-in locations may have few limitations. For instance, the homes and work locations are not always labeled by the users. They can be labeled in an automated process based on the characteristics of the locations. For instance, check-ins at a residential building may label that location as a home; that place may not be the actual home of a user but could be his or her friend's home. Similarly, a place labeled as an entertainment location could be the work location of a user. In such cases, the results should be interpreted with caution. However, the location categories are very accurate for other activity categories. For example, a commercial building may have different types of shops including restaurants, bars, gyms, grocery stores, etc. Based on the names and crowd-sourced information, the check-in program can distinguish among these places and can categorize accurately.

3. Activity pattern model

3.1. Problem definition

Inferring individual activity patterns involves finding the complex multi-day patterns from an individual's everyday activity participation choices. Activity patterns can be represented as a distribution of activity labels where each activity label can be represented by the day of week, time and category of activity. The problem of inferring activity patterns is defined as: given that the activity labels for m weeks as $a_1^1, a_1^2, \ldots, a_1^{n_1}$; $a_2^1, a_2^2, \ldots, a_2^{n_2}$; $\ldots$; $a_m^1, a_m^2, \ldots, a_m^{n_m}$ where an individual participates in $n_1, n_2, \ldots, n_m$ activities on week $1, 2, \ldots, m$ respectively, determine the K latent activity patterns through $\phi_k, k \in 1, 2, \ldots, K$ where each activity pattern ϕ_k is a distribution of activity labels (see Fig. 2).

3.2. Background

In this section, we present a hierarchical mixture model of individual activity pattern inference. The particular model, called as topic model or Latent Dirichlet Allocation (LDA), was originally proposed by Blei et al. (2003) within the domain of machine learning. Topic models are generally used for finding the topics in a large corpus of documents where each document is modeled as a mixture of topics; each topic is modeled as a distribution of words where the entity "word" represents the basic unit of discrete data.

A topic model is a generative model specifying a probabilistic process for generating documents. This generative model is a collection of simple probabilistic sampling rules stating how words in a document can be generated based on latent variables (e.g., topics of a document). The procedure is as follows (Steyvers and Griffiths, 2007): to make a new document a distribution over topics is chosen first; next based on this distribution, for each word in the document, a topic is randomly chosen and a word is drawn from that topic. A statistical technique can be developed to reverse this process and find the set of topics that generate the collection of documents. The goal of fitting a generative model is thus to find the best set of latent variables (i.e. topics of the documents) that can explain the observed data (i.e. words in the documents).

The generative procedure involved in a topic model does not assume any particular order of the words in a document (Steyvers and Griffiths, 2007). The only relevant information to the model is the number of times each word appears in a document. This assumption is known as "bag-of-words" assumption in statistical models of language processing. Although the ordering of the words may have important information for predicting the content of a document, such information is not used by topic models. This assumption also holds for the proposed activity pattern models. Although determining the activity sequence is an important problem for activity-based modeling, the data set used in this analysis does not contain the complete sequence of the activities participated by an individual. Due to this limitation, ignoring the order of the activities would be a reasonable assumption. However, this assumption does not prevent us to consider the temporal dimensions of activity participation. We can represent the timing of activity participation in the model.

A topic model can be applied to other types of discrete data also. For activity recognition, it was first applied to wearable sensor data by Huynh et al. (2008). Later Farrahi et al. (2011) applied it for mobile phone location-sequence data.

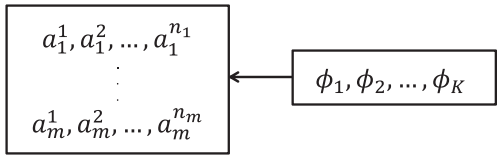


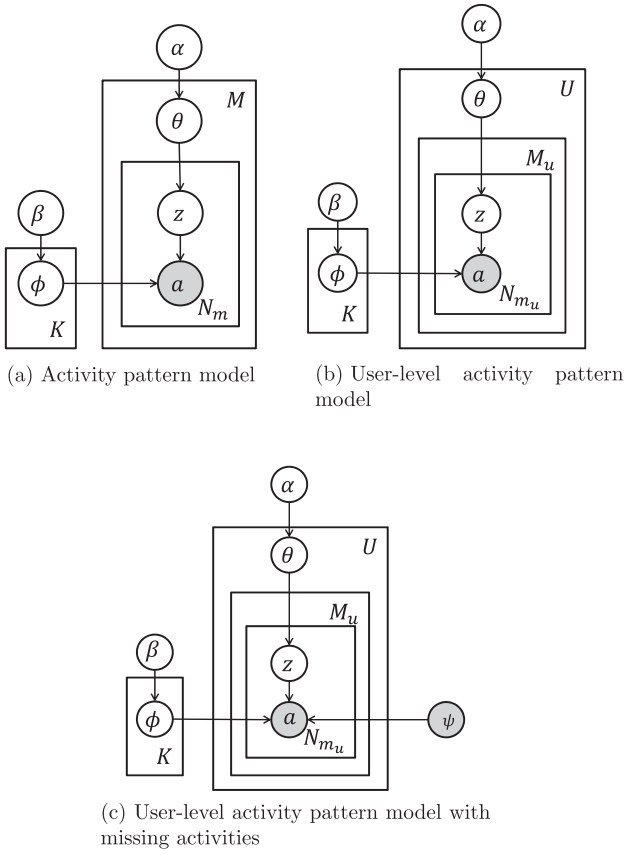
Fig. 2. Activity pattern inference problem.

They replaced words with location-sequences, documents with days and topics with routines. Their model produced the probability of a location-sequence for each routine and the probability of a routine for each day. However, they applied the model for mobile phone data which retains the sequential nature of activity participation. Unfortunately, the "bag-of-words" assumption of topic modeling ignores the ordering of activity participation. Thus a richer model can be developed from this location sequence data. On the other hand, social media data used in this study does not record the complete sequence of activity participation leading to a perfect application of topic modeling. Furthermore, they could consider only a limited number of activities such as activities participated at home, work and other places whereas we consider more number of activity categories providing a better understanding of urban activity patterns. In addition, previous work only modeled daily activity patterns whereas we develop a multi-day (i.e., weekly) activity pattern model.

3.3. Model description

In this section, we present a brief description of the proposed topic model. An activity label is defined as the basic unit of data to be picked from a set of possible activity labels of size A; a user participates in N activities in a week and the user has a collection of activities for M weeks; and a pattern represents the collection of activity participation and there are K latent patterns in user behavior, where K is defined by the analyst.

The probabilistic generative process in a topic model of activity pattern recognition starts by choosing a distribution over patterns $\mathbf{z} = (z_{1:K})$ for a given week. Given a distribution of patterns for a week, activity labels are generated by assigning a pattern from this distribution. The result is a vector of n_m activity labels $\mathbf{a} = (a_{1:n_m})$ for a week m. Pattern mixture parameters θ and ϕ are assumed to have Dirichlet prior distributions; where θ is an $M \times K$ matrix of week-specific mixture weights for the K patterns, each drawn from a Dirichlet prior, with hyper-parameter α and ϕ is a $A \times K$ matrix of activity-specific mixture weights over A items for the K patterns, each drawn from a Dirichlet (β) prior, with hyper-parameter β. The generative process (Fig. 3a) can be summarized as:



(a) Activity pattern model

(b) User-level activity pattern model

(c) User-level activity pattern model with missing activities

Fig. 3. Topic models of activity pattern inference. White circles represent random variables, shaded circles represent observed variables, rectangles represent the repetitiveness of the data, and arrows represent the dependency between the entities.

1. For each activity pattern $k \in 1, 2, \ldots, K$ select a distribution over activity labels $\phi^{(k)} \sim Dirichlet(\beta)$.
2. For each week $m \in 1, 2, \ldots, M$.
   (a) Select a distribution over activity patterns $\theta^{(m)} \sim Dirichlet(\alpha)$.
   (b) For each activity label i in week m
      i. Select a pattern $z_i \sim Multinomial(\theta^{(m)})$; $z_i \in 1, 2, \ldots, K$.
      ii. From activity pattern z_i, select an activity label $a_i \sim Multinomial(\phi^{(z_i)})$; $a_i \in 1, 2, \ldots, A$.

Given M weeks of activities, K activity patterns over A unique activity labels, the main objectives of the inference of activity pattern classifications are:

1. To find the probability of an activity label a given each pattern k, $P(a|z = k) = \phi_k^a$, where $P(a|z = k)$ is represented with K multinomial distributions ϕ over activity labels of size A.
2. To find the probability of a pattern k for an activity label in week m, $P(z = k|m) = \theta_m^k$, where $P(z|m)$ is represented with M multinomial distributions θ over K activity patterns.

The above model views a weekly activity pattern as a probability distribution over activity labels and weekly activities as a mixture of these patterns. From K weekly activity patterns, the probability of ith activity in a given week m is:

$$P(a_i|m) = \sum_{j=1}^{K} P(a_i|z_i = j)P(z_i = j|m) \tag{1}$$

where z_i is the latent variable indicating the pattern from which the ith activity is drawn, $P(a_i|z_i = j)$ is the probability of the activity a_i under the jth pattern and $P(z_i = j|m)$ is the probability of choosing an activity from pattern j in the activities of week m. Intuitively, $P(a|z)$ determines the importance of an activity label to a pattern and $P(z|m)$ determines the prevalence of the patterns in weekly activities.

The complete model of the weekly activity generation is:

$$a_i|z_i, \phi^{(z_i)} \sim Multinomial(\phi^{(z_i)})$$
$$\phi \sim Dirichlet(\beta)$$
$$z_i|\theta^{(m)} \sim Multinomial(\theta^{(m)})$$
$$\theta \sim Dirichlet(\alpha)$$

where α and β are hyper parameters for the prior distributions of θ and ϕ respectively. We assume Dirichlet prior distributions which are conjugate to the multinomial distributions.

3.4. Parameter estimation

In practice, there are various approximate techniques for estimating the parameters of this model (Blei et al., 2003; Griffiths and Steyvers, 2004). In this work, we use the Gibbs sampling approach proposed by Griffiths and Steyvers (2004). The algorithm can be found in detail in Griffiths and Steyvers (2004). Only a brief description of the approach is provided here.

The estimation approach starts with the joint distribution of activity labels and patterns $P(\mathbf{a}, \mathbf{z})$ written as:

$$P(\mathbf{a}, \mathbf{z}) = P(\mathbf{a}|\mathbf{z})P(\mathbf{z}) \tag{2}$$

The first term is written as:

$$P(\mathbf{a}|\mathbf{z}) = \left(\frac{\Gamma(A\beta)}{\Gamma(\beta)^A}\right)^K \prod_{k=1}^{K} \frac{\prod_a \Gamma(n_k^a + \beta)}{\Gamma(n_k^{(\cdot)} + A\beta)} \tag{3}$$

where n_k^a is the number of times activity label a is assigned to pattern k and $n_k^{(\cdot)} = \sum_{a=1}^{A} n_k^a$.

And the second term is written as:

$$P(\mathbf{z}) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^M \prod_{m=1}^{M} \frac{\prod_k \Gamma(n_m^k + \alpha)}{\Gamma(n_m^{(\cdot)} + K\alpha)} \tag{4}$$

where n_m^k is the number of times an activity label from week m is assigned to pattern k and $n_m^{(\cdot)} = \sum_{k=1}^{K} n_m^k$.

To estimate the model parameters, a Markov Chain Monte Carlo (MCMC) procedure is used. In MCMC, samples are taken from a Markov chain constructed to converge to a target distribution. In our model, each state of the chain is the assignment of a pattern to an activity label and the transition from one state to another follows a specific rule based on Gibbs sampling approach (Robert, 2001). In this procedure, the next state is reached by sampling the variables from a conditional distribution which specifies the distribution of the variables conditioned on the current assignment of all other variables and the observations. To apply the Gibbs sampling approach, a pattern can be assigned to an activity label using the following conditional distribution (Griffiths and Steyvers, 2004):

$$P(z_i = k|\mathbf{z}_{-i}, \mathbf{a}) \propto \frac{n^{a_i}_{-i,k} + \beta}{n^{(\cdot)}_{-i,k} + A\beta} \frac{n^{k}_{-i,m_i} + \alpha}{n^{(\cdot)}_{-i,m_i} + K\alpha} \tag{5}$$

where n_{-i} is the count excluding the current pattern assignment of z_i. Eq. (5) expresses two ratios; the first ratio expresses the probability of activity label a_i in pattern k and the second ratio expresses the probability of pattern k in the activities of week m.

Finally, the parameters representing the hidden patterns can be computed as:

$$\hat{\phi}^{a}_{k} = \frac{n^{a}_{k} + \beta}{n^{(\cdot)}_{k} + A\beta} \tag{6}$$

$$\hat{\theta}^{k}_{m} = \frac{n^{k}_{m} + \alpha}{n^{(\cdot)}_{m} + K\alpha} \tag{7}$$

We developed a code in Python programming language to process the data and estimate the model parameters. To reduce the computational time significantly, we used an extension of Python called as Cython to write the core computational steps of Gibbs sampling procedure. The actual time required to estimate the parameters depends on the number of input activity labels, the number of latent patterns and the number of samples. A typical setup for the input data used in the paper took less than 30 min to estimate the parameters.

3.5. Applying activity pattern model to geo-location data

We represent an activity label with three identifiers which include: day of week, hour of activity and the type of activity. Thus an activity label "Mon13Ea" represents an eating out activity performed at Monday 1 pm. We combine each individual's weekly activities and run the Gibbs sampling algorithm (Griffiths and Steyvers, 2004) over all the users' activities to infer the latent patterns of weekly activities. We assume $\beta = 0.1$ and $\alpha = 50/K$ which are reasonable choices for topic models. For model selection, we run our algorithm for different number of activity patterns (K) and compute *perplexity* – a metric to measure how well the model can predict the unseen data for each run. We then select the optimal number of patterns based on perplexity values.

Perplexity is a standard metric in machine learning to measure the performance of a probabilistic model. It represents a function of the average likelihood of obtaining a test data set given a set of model parameters. Perplexity is defined as the exponential of the negative of average predictive likelihood of a test data given a model (Griffiths and Steyvers, 2004). Perplexity of a test data set, a set of activity labels $\{\mathbf{a_m}\}$ for $m \in D^{test}$, given a model $\mathcal{M}$ is defined as

$$Perplexity = \exp\left[-\frac{\sum_{m=1}^{M} \log p(\mathbf{a_m}|\mathcal{M})}{\sum N_m}\right] \tag{8}$$

where N_m is the number of activity labels in each weekly activities of week m and $p(\mathbf{a_m}|\mathcal{M})$ can be computed using Eq. (1). For estimating perplexity values we randomly split the data set with 90% of the users in the training set and rest 10% in the test set; estimate the activity pattern model on the training data set; and compute the perplexity values on the test data set. A lower perplexity value indicates a better performance of the model. It is found that with the increase of the number of activity patterns the perplexity value reduces; however there is no significant improvement beyond a certain number of patterns (Fig. 4).

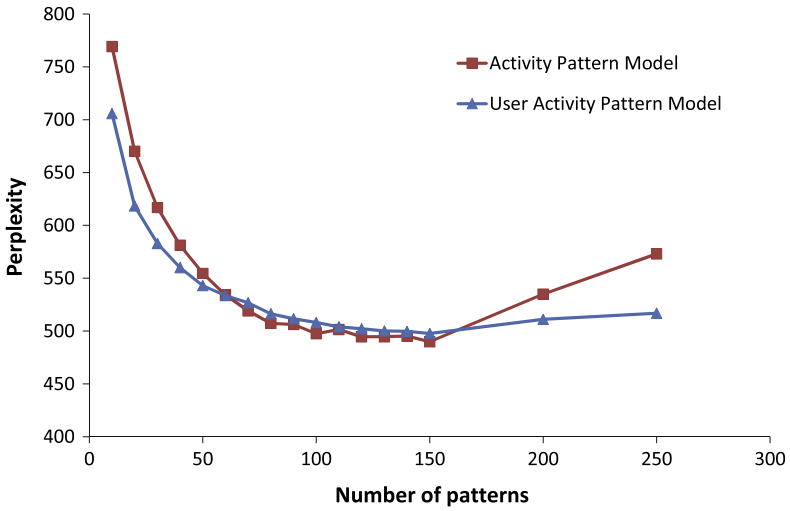


Fig. 4. Perplexity versus the number of activity patterns for activity pattern model and user activity pattern model.

Based on perplexity values, we select $K = 100$ for running the Gibbs algorithm for finding weekly activity patterns. Table 3 presents the results of the activity pattern model applied to the user activity data. Although we estimate 100 latent activity patterns, we report here a few interesting ones and the probabilities of the top 10 labels for each activity pattern. Few interesting activity patterns are as follows (Fig. 5):

- Pattern 1 captures the afternoon eating activities on weekdays.
- Pattern 2 captures the evening entertainment activities on Fridays and weekends (Saturday and Sunday).
- Pattern 3 captures the work-related activities around 1 pm on weekdays.

Table 3
Activity pattern model results.

a	$P(a\|Z)$	a	$P(a\|Z)$	a	$P(a\|Z)$	a	$P(a\|Z)$
Pattern1		**Pattern2**		**Pattern3**		**Pattern4**	
Wed13Ea	0.170814	Fri22Ea	0.249387	Tue13Wo	0.154818	Sun3En	0.16457
Thu13Ea	0.159907	Fri23En	0.204941	Wed13Wo	0.152067	Sun2En	0.137095
Fri13Ea	0.15675	Sun2Ea	0.04121	Mon13Wo	0.149316	Sun4Ea	0.095882
Tue13Ea	0.149287	Sat0En	0.031079	Thu13Wo	0.144315	Sun4En	0.083081
Mon13Ea	0.104222	Sat18Re	0.019641	Fri13Wo	0.115557	Sun1En	0.069968
Mon12Ea	0.025001	Tue23Wo	0.015719	Wed14Wo	0.031034	Sun5Ea	0.063411
Wed14Ea	0.006918	Thu23Ea	0.014739	Fri14Wo	0.025282	Sun3Ea	0.060601
Mon14Ea	0.006918	Tue23Ea	0.014412	Tue14Wo	0.024782	Sun5En	0.036561
Fri5Ho	0.005482	Wed1En	0.014412	Thu14Wo	0.021531	Sun1Ea	0.035936
Fri16Ea	0.005195	Sat20En	0.012451	Wed12Wo	0.010278	Sun8Ho	0.00971
Pattern5		**Pattern9**		**Pattern10**		**Pattern11**	
Wed0En	0.214706	Sat2Ea	0.223208	Mon18Ea	0.209771	Sun0Ea	0.141129
Wed2En	0.124052	Sat0En	0.183612	Tue18Ea	0.203177	Sat20Re	0.128775
Wed1En	0.112405	Sat3En	0.067116	Wed18Ea	0.139439	Sat21Re	0.118372
Fri2En	0.108313	Sun1En	0.0501	Tue15Ea	0.040849	Sat23Re	0.074807
Thu1En	0.072744	Fri22So	0.042246	Thu16Ea	0.037709	Sat22Re	0.06668
Wed3En	0.033712	Fri23Ea	0.020976	Fri18Ea	0.023894	Fri23Re	0.042622
Sat1En	0.031509	Sun23Ea	0.01214	Wed16Ea	0.02044	Sat18Ea	0.03872
Wed22En	0.012308	Wed18Tr	0.008868	Wed19Ea	0.010393	Sat23Ea	0.032868
Mon1En	0.011993	Sat20Ea	0.008541	Sat18Ea	0.008195	Sun16Ea	0.015638
Wed1Ea	0.007586	Sun0En	0.008214	Sat2So	0.007567	Fri23Ea	0.012387
Pattern15		**Pattern18**		**Pattern24**		**Pattern27**	
Sat15Ea	0.192541	Sun18Ea	0.251463	Sun17Sh	0.161624	Thu0Ho	0.023998
Sat14Ea	0.133772	Sun19Sh	0.170815	Sun16Ea	0.153313	Wed1Ho	0.023177
Sun15Ea	0.08358	Sun19Ea	0.083526	Sun17Ea	0.123765	Tue0Ho	0.023013
Sat14Sh	0.068649	Sun18Sh	0.060755	Sun16Sh	0.099141	Wed0Ho	0.020877
Sun14Ea	0.049271	Sat21Ea	0.028496	Sun18Sh	0.083444	Tue23Ho	0.020549
Sun16Ea	0.041647	Sun20Sh	0.028496	Sun16So	0.046508	Tue2Ho	0.020385
Sat13Ot	0.033705	Sun16Ea	0.026914	Sun17So	0.034196	Fri1Ho	0.019728
Sat15Sh	0.030528	Sun20Ea	0.014264	Sun17Tr	0.013882	Thu1Ho	0.018906
Sat14So	0.029575	Sat3Ea	0.012366	Sat17Ea	0.011419	Thu3Ho	0.018742
Wed23Ea	0.027669	Sun18So	0.011417	Sun15So	0.011419	Mon0Ho	0.018414
Pattern40		**Pattern48**		**Pattern50**		**Pattern85**	
Sat20Sh	0.177484	Fri17Sh	0.10022	Thu12Ea	0.160143	Tue22En	0.132209
Sat20Ea	0.146347	Tue17Sh	0.085995	Tue12Ea	0.149629	Fri21En	0.101261
Sat19Ea	0.116418	Wed17Sh	0.079811	Wed12Ea	0.146625	Thu21En	0.097392
Sat19Sh	0.113698	Thu17Sh	0.07208	Fri12Ea	0.128901	Wed22En	0.087398
Sat21Sh	0.107651	Wed16Sh	0.062185	Mon12Ea	0.100964	Wed21En	0.078371
Sat21Ea	0.087397	Fri16Sh	0.059402	Tue16Ea	0.036979	Thu22En	0.069022
Sat22Sh	0.010006	Thu16Sh	0.051362	Thu16Ea	0.021058	Tue21En	0.064186
Sat19Tr	0.008193	Fri15Sh	0.045178	Mon16Ea	0.012647	Fri19En	0.027757
Mon1Ot	0.007588	Fri15Ea	0.031881	Mon12Wo	0.012346	Fri18En	0.020342
Thu2So	0.007286	Tue16Sh	0.030644	Mon13Ea	0.011746	Fri22En	0.016474
Pattern87		**Pattern89**		**Pattern98**		**Pattern100**	
Thu14Ea	0.101778	Sat0Ea	0.215027	Sat2En	0.23773	Thu18Ea	0.215151
Wed14Ea	0.095566	Fri0Ea	0.106736	Sat1En	0.217796	Fri17Ea	0.108252
Fri14Ea	0.094975	Sun21Ea	0.084188	Sat4En	0.096077	Tue17Ea	0.099013
Tue14Ea	0.09113	Sun22Ea	0.080695	Sat3En	0.089432	Wed17Ea	0.068
Thu15Ea	0.090834	Sun0Ea	0.079425	Sat0En	0.052886	Thu18Sh	0.062721
Mon15Ea	0.076932	Sun23Ea	0.067039	Fri23En	0.036878	Thu18So	0.032697
Wed15Ea	0.06865	Sun1Ea	0.054654	Sat5En	0.018756	Thu17Ea	0.028737
Mon14Ea	0.062143	Wed23En	0.020039	Sat1So	0.008789	Mon16Ea	0.022799
Tue15Ea	0.050608	Sat19Ea	0.012735	Thu22En	0.008487	Thu19Ea	0.020489
Fri15Ea	0.041439	Sun3Ea	0.010829	Sun17Ea	0.007279	Thu18Wo	0.01389

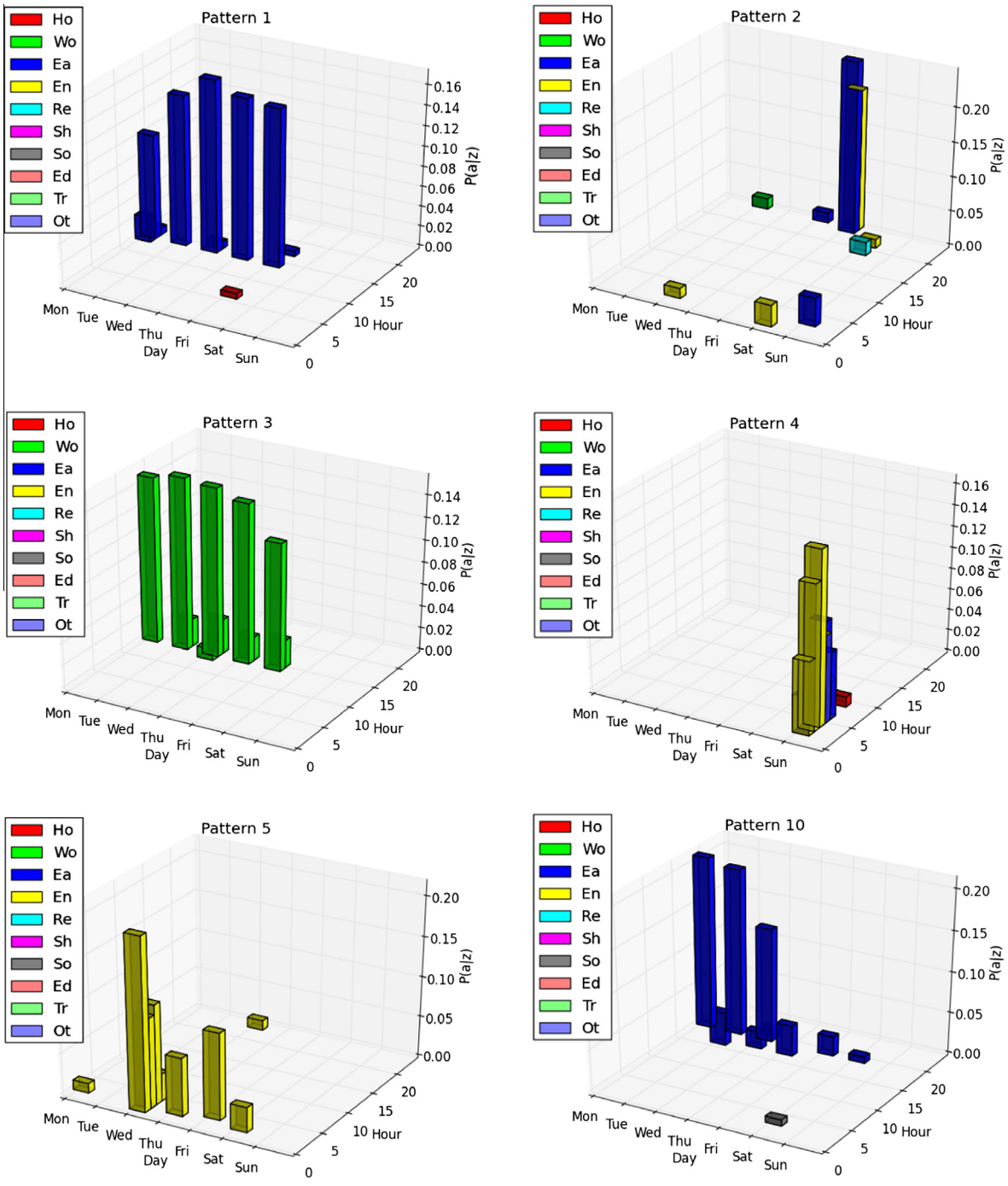


Fig. 5. Activity pattern distributions.

- Pattern 4 captures the late-night entertainment activities on Sundays.
- Pattern 5 captures the post-midnight entertainment activities on weekdays with Wednesday as the most likely day for this type of activities.
- Pattern 11 captures the eating and recreational activities on weekend evenings.
- Patterns 10 and 100 capture the eating activities on weekday evenings while pattern 15 captures the same for weekend evenings.
- Pattern 27 captures the at-home activities in a week; since the home-related activities are less shared in social media, there are no dominant activity labels in this pattern.
- Pattern 48 captures the evening shopping and eating activities on weekdays.

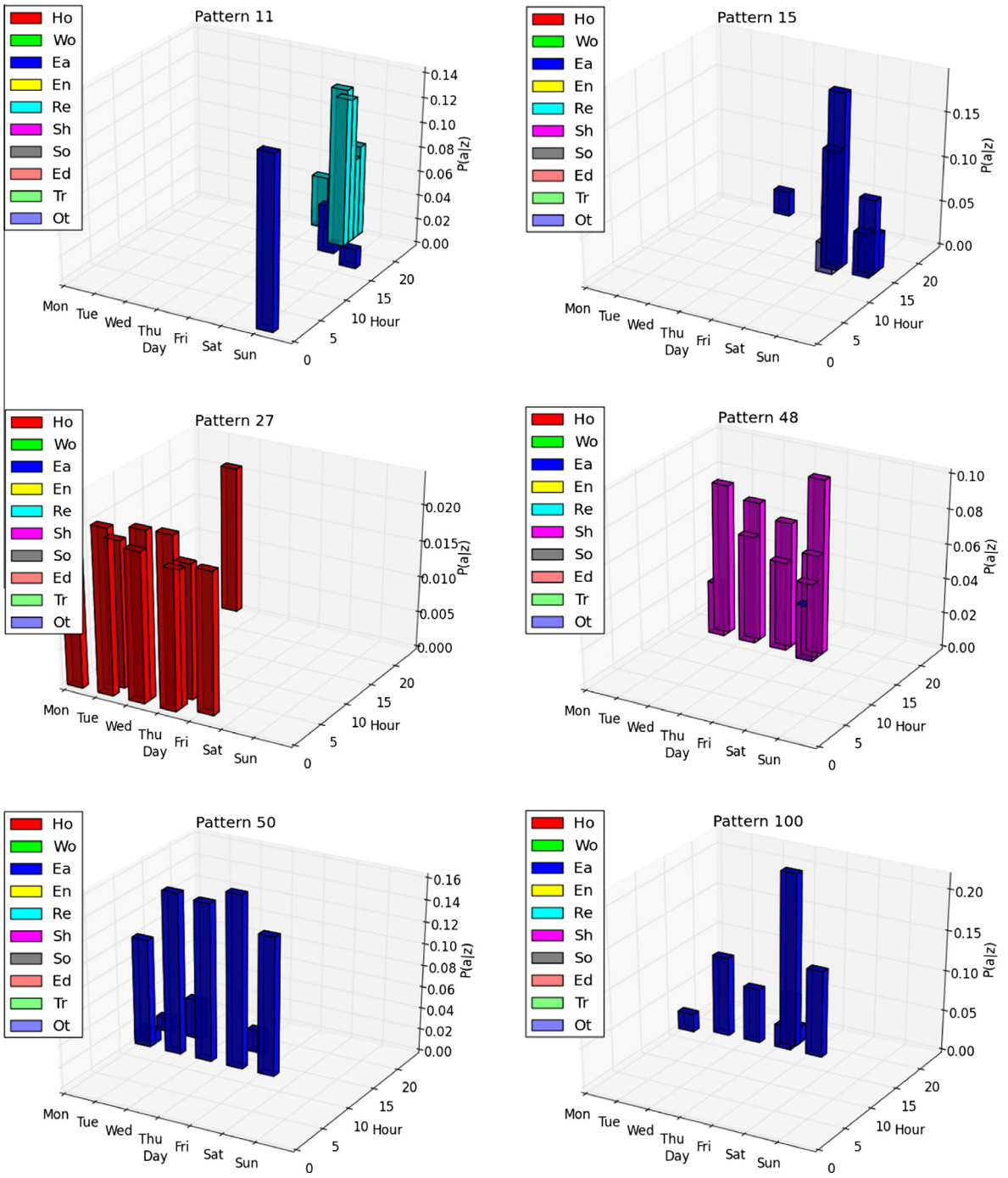


Fig. 5 (*continued*)

Despite the noise present in the data, the model results indicate that the proposed topic model of activity pattern is a promising method to find the latent patterns of individual weekly activities. The observed patterns mostly contain the non-work related activities with a higher concentration of the entertainment and eating activities. However, it is difficult to observe these flexible activities in traditional surveys based on an individual's typical working day activities. These activity patterns are plausible in terms of activity type, day of week and time of activity participation. Using these patterns, we can generate the typical non-work related activities of an individual for a week.

4. User activity pattern model

The activity pattern model (Section 3.3) when applied over all the users assumes that each week consists of activity participation of a unique user. It ignores the fact that a user may have activity participation over multiple weeks. Thus the model cannot consider user-level patterns. In this section, we extend the activity pattern model where user-specific patterns are represented.

4.1. Model description

The probabilistic generative process (Fig. 3b) for the model is summarized as:

1. For each activity pattern $k \in 1, 2, \ldots, K$ select a distribution over activity labels $\phi^{(k)} \sim Dirichlet(\beta)$.
2. For each user $u \in 1, 2, \ldots, U$ select a distribution over activity patterns $\theta^{(u)} \sim Dirichlet(\alpha)$.
   (a) For each week $m \in 1, 2, \ldots, M_u$ of user u.
      i. For each activity label i in week m
         A. Select a pattern $z_i \sim Multinomial(\theta^{(u)}); z_i \in 1, 2, \ldots, K$.
         B. From activity pattern z_i select an activity label $a_i \sim Multinomial(\phi^{(z_i)}); a_i \in 1, 2, \ldots, A$.

For estimation purpose, the second term (as written in Eq. (4)) can be written as:

$$P(\mathbf{z}) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^U \prod_{u=1}^{U} \frac{\prod_k \Gamma(n_u^k + \alpha)}{\Gamma(n_u^{(\cdot)} + K\alpha)} \tag{9}$$

where n_u^k is the number of times an activity label from user u is assigned to pattern k and $n_u^{(\cdot)} = \sum_{k=1}^{K} n_u^k$

For applying the Gibbs sampling approach, a pattern can be assigned using the following conditional distribution:

$$P(z_i = k | \mathbf{z}_{-i}, \mathbf{a}, u) \propto \frac{n_{-i,k}^{a_i} + \beta}{n_{-i,k}^{(\cdot)} + A\beta} \frac{n_{-i,u_i}^{k} + \alpha}{n_{-i,u_i}^{(\cdot)} + K\alpha} \tag{10}$$

where n_{-i} is the count excluding the current pattern assignment of z_i. Intuitively, Eq. (10) expresses two ratios; the first ratio expresses the probability of activity label a_i in pattern k and the second ratio expresses the probability of pattern k in the activities of user u. Finally, the parameters can be computed as:

$$\hat{\phi}_k^a = \frac{n_k^a + \beta}{n_k^{(\cdot)} + A\beta} \tag{11}$$

$$\hat{\theta}_u^k = \frac{n_u^k + \alpha}{n_u^{(\cdot)} + K\alpha} \tag{12}$$

4.2. Applying user activity pattern model to geo-location data

In this section, we present the results from user activity pattern models. Similar to activity pattern model selection, we use perplexity values to determine the number of activity patterns to be estimated. Fig. 4 shows the results for perplexity measurements for user activity pattern model. Based on these perplexity values we select $K = 100$ for estimating the model parameters. Table 4 presents a few of the activity patterns estimated and the probability of the top 10 activity labels for each pattern reported. Few interesting patterns found from the results are as follows (Fig. 6):

- Pattern 2 captures the evening eating activities on weekdays.
- Pattern 7 captures the work-related activities around noon on weekdays.
- Pattern 9 captures the social activities on weekday evenings.
- Pattern 12 captures the late-night entertainment activities on weekdays.
- Pattern 13 captures the education related activities on weekdays.
- Patterns 17, 21 and 25 capture the eating activities on weekdays and weekends.

Similar to the previous activity patterns (Section 3.5), these user-level patterns demonstrate the applicability of our modeling approach to extract the hidden patterns of user activities. The reported patterns contain the likely activities participated by the users on different days of a week and hours of a day. We also report the top users for each of the activity patterns along with their corresponding probabilities. These users indicate the top contributors for the corresponding patterns. To demonstrate how well our methodology can extract the weekly activity patterns from a user's geo-location information, we report the distribution of activity labels of a user, his top most pattern proportions ($P(z|u) > 0.05$) and the

Table 4
User activity pattern model results.

Pattern2		**Pattern7**		**Pattern9**		**Pattern12**	
a	$P(a\|Z)$	a	$P(a\|Z)$	a	$P(a\|Z)$	a	$P(a\|Z)$
Thu19Ea	0.112754	Tue12Wo	0.109849	Wed19So	0.039626	Thu0En	0.143275
Mon19Ea	0.080108	Mon12Wo	0.097542	Mon20So	0.039031	Thu1En	0.110448
Wed19Ea	0.079829	Wed12Wo	0.094702	Thu20So	0.03784	Wed23En	0.087171
Thu20Ea	0.070063	Thu12Wo	0.092177	Fri18So	0.037542	Fri1En	0.065385
Fri19Ea	0.065041	Fri12Wo	0.073874	Wed17So	0.036053	Mon23En	0.056433
Tue19Ea	0.061414	Thu11Wo	0.030957	Tue20So	0.034565	Tue0En	0.056134
Wed20Ea	0.055833	Tue13Wo	0.03001	Fri19So	0.034267	Fri0En	0.045689
Fri20Ea	0.055554	Mon13Wo	0.029064	Wed18So	0.033076	Fri2En	0.042108
Mon20Ea	0.050811	Fri13Wo	0.028433	Fri16So	0.032481	Thu2En	0.04181
Tue20Ea	0.035185	Mon11Wo	0.028433	Fri20So	0.030695	Wed1En	0.041511
u	$P(u\|Z)$	u	$P(u\|Z)$	u	$P(u\|Z)$	u	$P(u\|Z)$
16049560	0.034569	37658109	0.026539	33549238	0.023108	21171366	0.010595
96633510	0.013415	17290220	0.026238	17218135	0.015415	104077681	0.007739
13823722	0.010202	25671564	0.023226	28068128	0.01513	15678504	0.006311
102202457	0.008328	15354898	0.015996	39551086	0.010571	16123855	0.006026
11282572	0.008328	14599416	0.015092	102202457	0.009716	15831495	0.006026
14217340	0.007256	16808381	0.015092	14080008	0.009146	14492779	0.006026
14182453	0.006453	96794283	0.015092	104077681	0.009146	24100009	0.00574
24440172	0.006185	131391305	0.014188	17388259	0.008292	23133239	0.005455
14774729	0.005918	803705	0.013586	24104668	0.007722	15811158	0.005455
22735032	0.00565	14573910	0.013285	17828046	0.007152	14246373	0.005169
Pattern13		**Pattern17**		**Pattern21**		**Pattern25**	
a	$P(a\|Z)$	a	$P(a\|Z)$	a	$P(a\|Z)$	a	$P(a\|Z)$
Tue13Ed	0.04975	Thu13Ea	0.148742	Sun19Ea	0.174563	Sat16Ea	0.145284
Wed13Ed	0.041254	Fri13Ea	0.132748	Sat18Ea	0.139791	Sun15Ea	0.12097
Tue12Ed	0.039051	Tue13Ea	0.119841	Sat17Ea	0.085961	Sun16Ea	0.102024
Mon13Ed	0.037478	Mon13Ea	0.091501	Sat19Ea	0.056873	Sat17Ea	0.050554
Fri13Ed	0.036534	Wed14Ea	0.04829	Sun17Ea	0.041493	Sat14Ea	0.049607
Mon14Ed	0.036534	Thu14Ea	0.03903	Fri18Ea	0.027116	Sun14Ea	0.028135
Thu12Ed	0.034331	Mon14Ea	0.028648	Thu23Ea	0.026447	Sat0Ea	0.025293
Tue14Ed	0.032443	Fri14Ea	0.026965	Sat16Ea	0.025444	Fri16Ea	0.022767
Thu14Ed	0.031499	Tue14Ea	0.023037	Sat0Ea	0.019091	Sat16Sh	0.021819
u	$P(u\|Z)$	u	$P(u\|Z)$	u	$P(u\|Z)$	u	$P(u\|Z)$
30481498	0.030073	9863222	0.021564	14571619	0.012128	14213141	0.017211
17044112	0.027369	27796272	0.018064	96633510	0.011173	16585201	0.012087
124308038	0.021661	136144242	0.013757	21003752	0.007035	96633510	0.010881
15712496	0.018957	35238243	0.011872	21970561	0.005761	19625559	0.009073
6095002	0.014751	134863077	0.011064	15530973	0.005761	7344992	0.007867
14688415	0.01415	21902908	0.011064	14052720	0.005443	19185772	0.007566
36302100	0.01385	9905762	0.010257	34083092	0.005125	14466465	0.006963
14467582	0.013549	6611042	0.010257	21162214	0.004806	85218012	0.00636
61272186	0.012648	21003752	0.009449	27270877	0.00417	34997518	0.00636
17283207	0.012047	14276033	0.007834	27013777	0.00417	7211482	0.006059

corresponding pattern descriptions (Fig. 7). This example demonstrates how a user's weekly activities can be modeled as a mixture of patterns as the user has participated in several activities that belong to few specific patterns. The user also repeats some activities weekly; for instance, the user has participated in entertainment activities close to midnight on several Wednesdays.

In addition to the activity label proportions, the user pattern model can determine user-specific pattern proportions ($\hat{\theta}_u^k$). These pattern proportions can be used to find the similarity among the users based on their activity participation behavior.

5. User activity pattern model with missing activities

One of the major limitations of social media data sets is the lack of some activities being reported by the users. Most likely these activities include activity participation at home and work places. Although the previous models can infer the weekly patterns of secondary activities, they cannot predict the complete activity routines that include both fixed and secondary activities. In this section, we extend the user-pattern model to incorporate the missing activities. The idea here is to enrich the model with background knowledge of individual activity participation (Steyvers, 2010). Here we include additional activity patterns (called as activity features) including fixed and secondary activities. The proposed model draws an activity label from the features in addition to the hidden activity patterns.

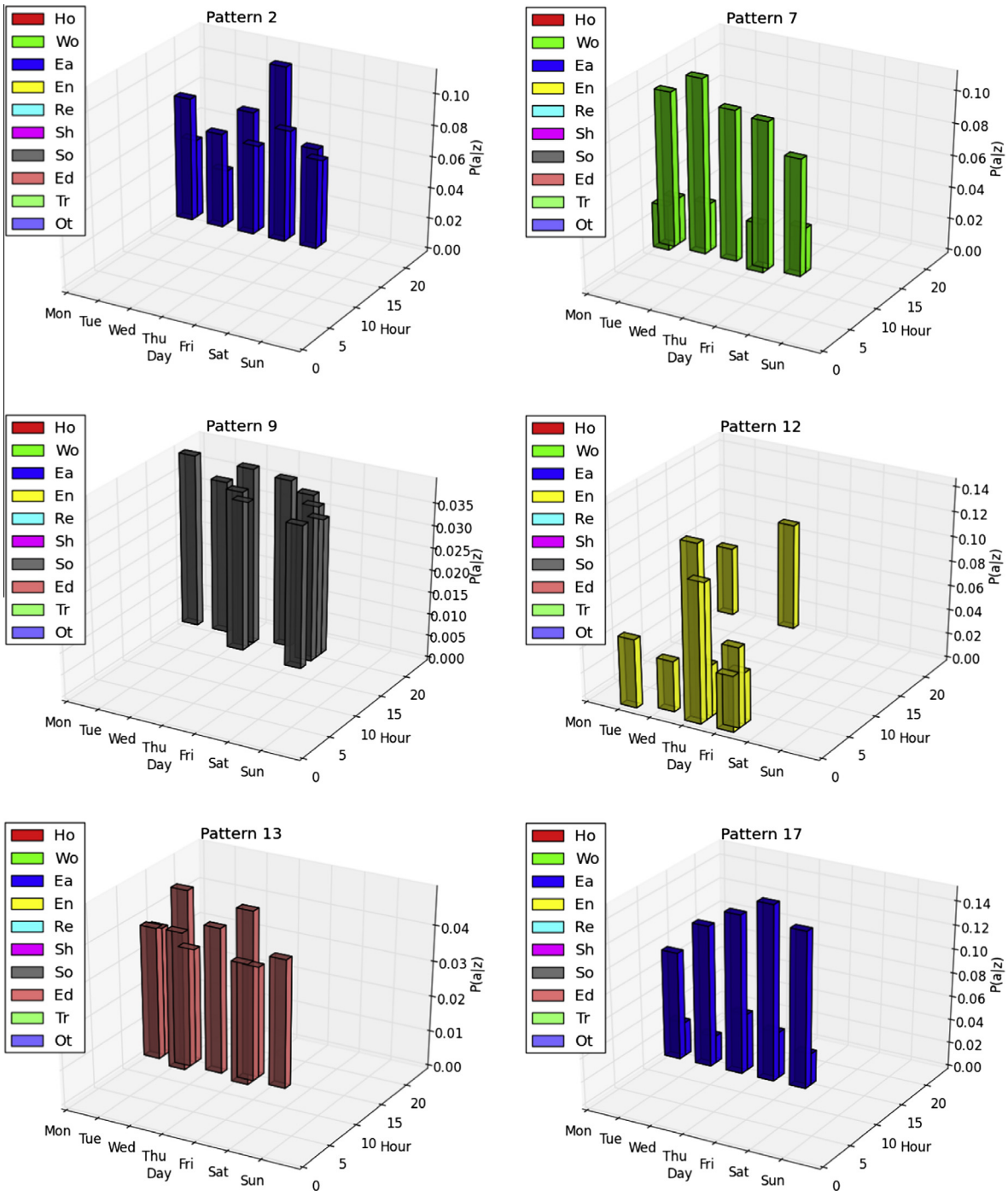


Fig. 6. User activity pattern distributions.

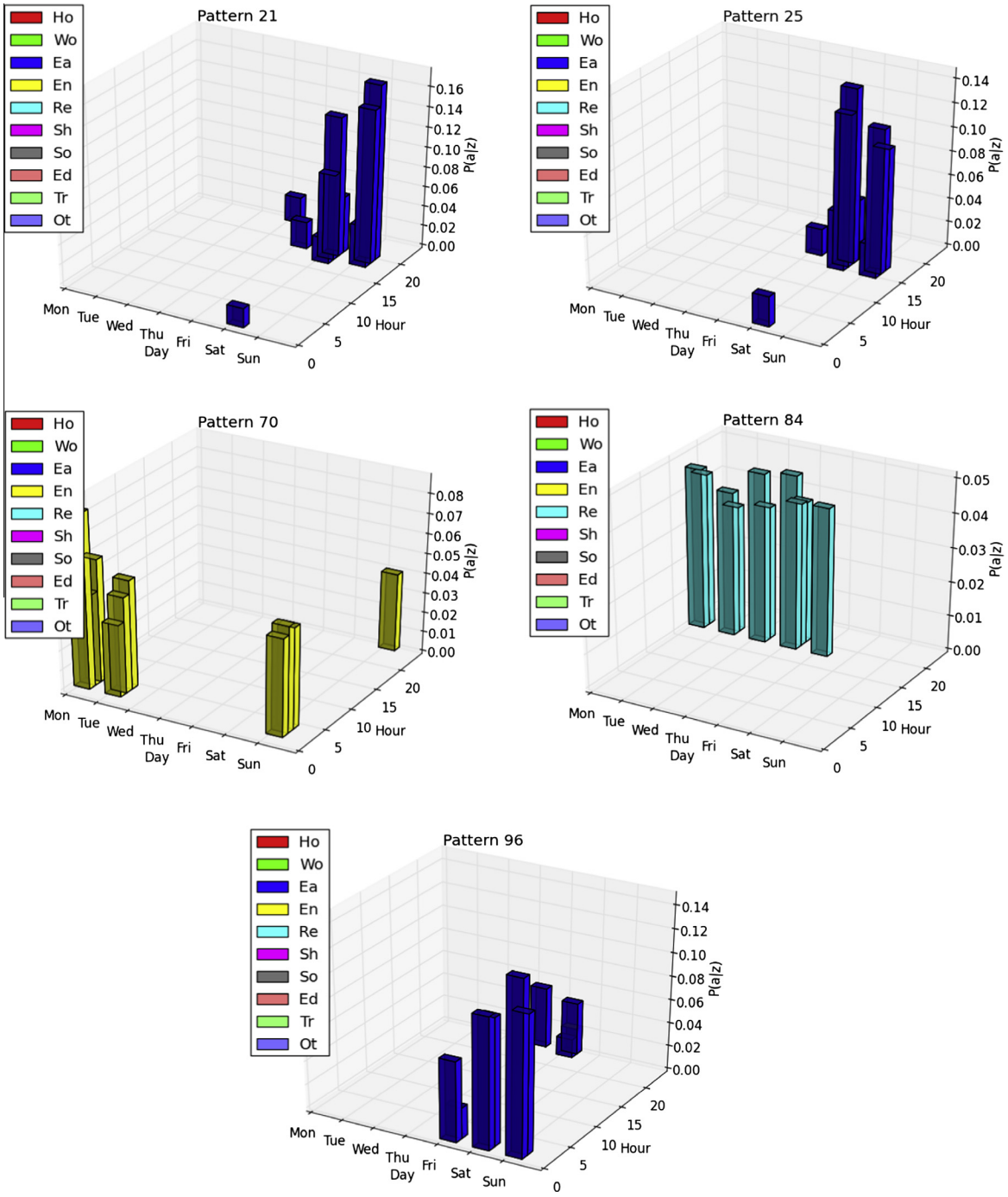


Fig. 6 (*continued*)

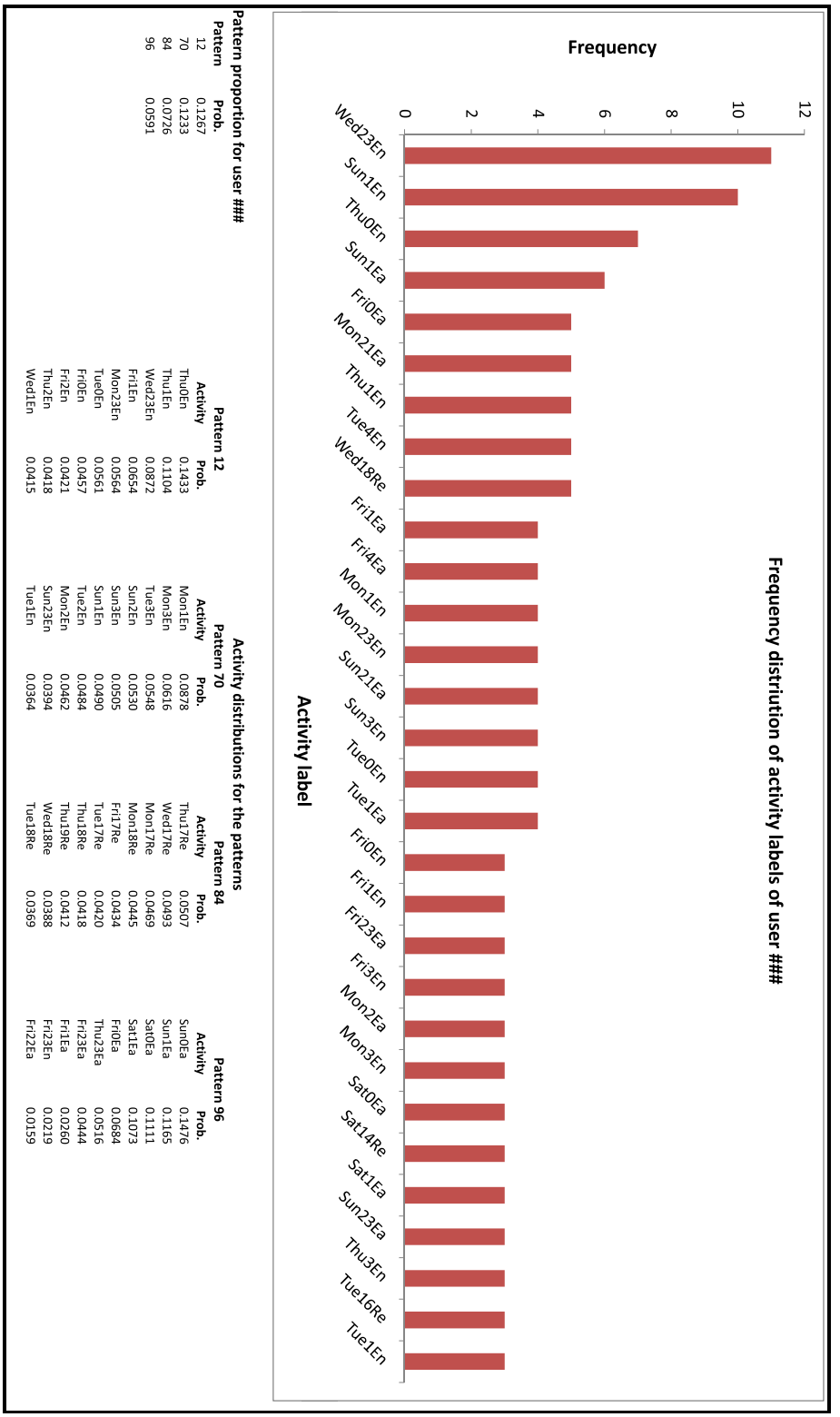


Fig. 7. The distribution of activity labels of a user, the pattern proportion and the activity distributions for the corresponding patterns. For activity label distribution, only the labels with frequencies not less than three are presented; the data consists of 19 weeks of activity participation of the user.

5.1. Model description

The probabilistic generative process (Fig. 3c) for the model is summarized as:

1. For each activity pattern $k \in 1,2,\ldots,K$ select a distribution over activity labels $\phi^{(k)} \sim Dirichlet(\beta)$.
2. For each activity feature $f \in 1,2,\ldots,F$ associate a predefined activity label distribution $\psi^{(f)}$.
3. For each user $u \in 1,2,\ldots,U$ select a distribution over activity patterns $\theta^{(u)} \sim Dirichlet(\alpha)$.
   (a) For each week $m \in 1,2,\ldots,M_u$ of user u.
      (i) For each activity label i in week m.
         A. Select a pattern $z_i \sim Multinomial(\theta^{(u)}); z_i \in 1,2,\ldots,K$.
         B. If $z_i \leqslant K$, from activity pattern z_i, select an activity label $a_i \sim Multinomial(\phi^{(z_i)}); a_i \in 1,2,\ldots,A$. Otherwise, from feature $f = z_i - K$, select an activity label $a_i \sim Multinomial(\psi^{(f)}); a_i \in 1,2,\ldots,A$.

For estimation purpose, the second term written in Eq. (4) can be written as:

$$P(\mathbf{z}) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^U \prod_{u=1}^{U} \frac{\prod_k \Gamma(n_u^k + \alpha)}{\Gamma(n_u^{(\cdot)} + K\alpha)} \tag{13}$$

where n_u^k is the number of times an activity label from user u is assigned to pattern k and $n_u^{(\cdot)} = \sum_{k=1}^{K} n_u^k$.

To apply the Gibbs sampling approach, a pattern can be assigned using the following conditional distribution: if $1 \leqslant z_i \leqslant K$

$$P(z_i = k|\mathbf{z}_{-i}, \mathbf{a}, u) \propto \frac{n_{-i,k}^{a_i} + \beta}{n_{-i,k}^{(\cdot)} + A\beta} \frac{n_{-i,u_i}^{k} + \alpha}{n_{-i,u_i}^{(\cdot)} + (K+F)\alpha} \tag{14}$$

if $z_i = K + f$ and $1 \leqslant f \leqslant F$

$$P(z_i = f|\mathbf{z}_{-i}, \mathbf{a}, u) \propto \frac{n_{-i,f}^{a_i} + \beta}{n_{-i,f}^{(\cdot)} + A\beta} \frac{n_{-i,u_i}^{f} + \alpha}{n_{-i,u_i}^{(\cdot)} + (K+F)\alpha} \tag{15}$$

where n_{-i} is the count excluding the current pattern assignment of z_i. Finally model parameters can be computed as:

$$\hat{\phi}_k^a = \frac{n_k^a + \beta}{n_k^{(\cdot)} + A\beta} \tag{16}$$

$$\hat{\theta}_u^k = \frac{n_u^k + \alpha}{n_u^{(\cdot)} + K\alpha} \tag{17}$$

5.2. Applying user activity pattern model with missing activities to geo-location data

To apply the user activity pattern model with missing activities, first a distribution of the activity patterns including the missing activities are required. These distributions can be found from traditional activity diary data that includes all the activities. Running the user activity pattern model on this activity diary data would give a distribution of those activity features. Since we do not have such activity diary data compatible with the current data set, we create an activity feature set from our data. We first split our original data set into two parts (i) training data set (70% of the users) and (ii) testing data set (30% of the users). Assuming that the training data set has no missing activities we run our user activity pattern model on this training data set. Activity patterns obtained from this model constitute the activity features. For creating users' activity sequences with missing activities, we intentionally remove few activities from our testing data set. For this experiment, all the activities on Tuesdays are removed from the testing data set. With features from the training data set, we then run the model described in Section 5.1 on this testing data set with missing activities. For running this model we assume that $K = 50,\ \beta = 0.1$ and $\alpha = 50/\mathrm{K}$.

One important task of the activity pattern model is to predict the missing activities. The probability of an activity label can be computed using the following equation:

$$P(a_i|u) = \sum_{j=1}^{K+F} P(a_i|z_i = j)P(z_i = j|u) \tag{18}$$

Using Eq. (18), we compute the predictive probabilities of all the activities that are not present in the weekly activities of an individual in the testing data set. Based on these predictive probabilities, we rank all the missing activity labels and express them as percentile of the total number of activity labels. The average of the ranking of only the missing activities are then computed. We run experiments for different number of features obtained from the training data set and report the corresponding average ranking of the missing activities (Fig. 8). With the increase of the features from the training data

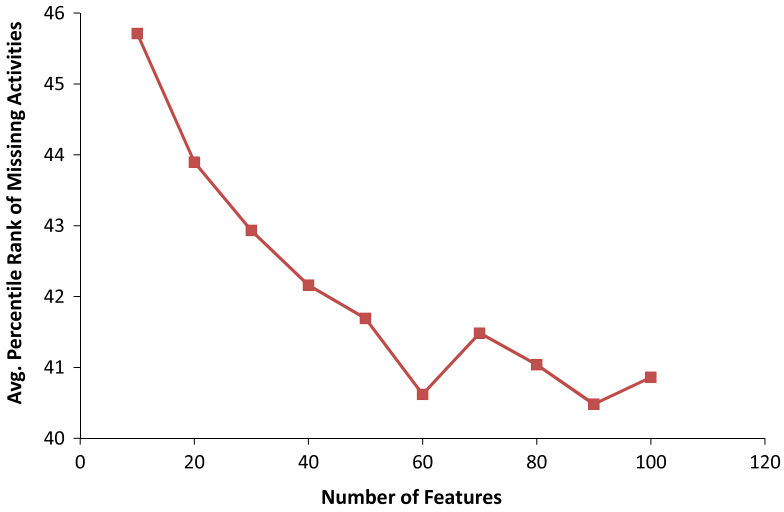


Fig. 8. Average percentile rank of the missing activities versus the number of features.

set the predictability of missing activities improves (lower the rank higher the probability of having the activity in the week). This shows the importance of these additional features in predicting the missing activities.

Our analysis regarding the missing activities is exploratory. We can only demonstrate the improvement in the predictions of the missing activities with the increase of the number of activity features. However several other questions still remain. How well the model can predict the missing activities? What are the roles of the proportions of the missing activities in the data for predicting activities? How can we determine an ideal set of activity features? How can we build a model that includes traditional activity diaries and geo-location data and infer activity choice patterns? We believe that finding the answers to these questions are beyond the scope of this paper. However future research is needed on these issues related to missing activities.

6. Conclusions

We have demonstrated the uses of geo-location data from social media for classifying individual activities. We propose a topic model to extract multi-day patterns of individual activities. The model views an activity pattern as a multinomial distribution of activity labels and individual activities as a mixture of activity patterns. We find several weekly activity patterns from individual activity information shared in social media. Observed patterns mainly capture an individual's participation in the non-work related flexible activities. We extend this model to capture user-specific patterns and identify the top users that contribute to a specific pattern. Finally, we extend this model to account for missing activities. We demonstrate that with additional information the model can predict well the missing activities in individual activity behavior.

The proposed models hold tremendous potential for activity behavior analysis as more and more geo-location data will be available in future. With the widespread uses of smartphones and check-in services in online social media, our approach can become useful in activity analysis. The developed models can classify and recognize urban activity patterns; such a classification can act as a basis for further empirical analysis based on geo-location data. The observed patterns can be used to classify individuals whose socio-demographic features are not available from social media. Given the potential availability of these data sets in near future and the lack of appropriate methodologies to characterize this information, the proposed methods can significantly contribute towards understanding human activity choice behavior.

However, our analysis has several limitations which include:

1. The derived activity patterns do not have sufficient explanatory power. In other words, we cannot explain why different individuals have different distributions of activity patterns.
2. We do not observe the complete sequence of the activities participated by an individual. This lack of activity sequences poses a major drawback against predicting the activity routines of an individual.
3. One major limitation of our data set is the potential lack of representativeness of the population behavior. It is found that users of smartphones and location-based services have minor over representation from younger people (Comscore, 2011).

It is important to note that these limitations are related to the unique features of the data set used in the analysis and not the methodological approach that we propose. A few of these limitations may go away with wider uses of social media over the population. For instance, it is expected that with the wide-spread uses of location-based services in the future, sample representativeness of social media data sets will improve. None the less, given the available data we are able to extract the user patterns and we anticipate that the observed patterns are not the exclusive traits of the social media users.

Some of these limitations however will require further research. It is possible to provide behavioral underpinnings to different patterns observed from social media data. From geo-location data, information on individual attitudes or interests to different activity locations and urban neighborhoods can be gathered and individual life-style patterns can be constructed (Hasan, 2013). Such an analysis will enable us to cluster the users based on their life-style choices and correlate activity patterns with those choices. This life-style concept may provide a richer medium to explain different behavioral patterns observed from social media data sets.

Another potential future research direction will be to reconstruct individual activity sequences from incomplete information on activity participation. A predictive model can be developed for the daily sequences of individual activities using the dependencies among activity types, locations and durations. Such a model can reconstruct user mobility trajectories and activity sequences based on the partial information available in geo-location data from social media.

References

Alesiani, Francesco, Gkiotsalitis, Konstantinos, Baldessari, Roberto, 2014. A probabilistic activity model for predicting the mobility patterns of homogeneous social groups based on social network data. In: The 93rd Annual Meeting of Transportation Research Board.

Allahviranloo, Mahdieh, Recker, Will, 2013. Daily activity pattern recognition by using support vector machines with multiple classes. Transport. Res. Part B: Methodol. 58, 16–43.

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

Cebelak, Meredith Kimberly, 2013. Location-based social networking data: doubly-constrained gravity model origin-destination estimation of the urban travel demand for Austin, TX. Master's Thesis, The University of Texas at Austin.

Cheng, Zhiyuan, Caverlee, James, Lee, Kyumin, Sui, Daniel Z., 2011. Exploring millions of footprints in location sharing services. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM).

Collins, Craig, Hasan, Samiul, Ukkusuri, Satish V., 2013. A novel transit rider satisfaction metric: rider sentiments measured from online social media data. J. Public Transport. 16 (2).

Comscore, 2011. Comscore Press Release May 12, 2011. <http://www.comscore. com/Press_Events/Press_Releases/2011/5/Nearly_1_in_5_Smartphone_Owners_A ccess_Check-In_Services_Via_their_Mobile_Device> (accessed 12.01.12).

Cranshaw, Justin, Schwartz, Raz, Hong, Jason I, Sadeh, Norman M., 2012. The livehoods project: utilizing social media to understand the dynamics of a city. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM).

De Choudhury, Munmun, Gamon, Michael, Counts, Scott, Horvitz, Eric. 2013. Predicting depression via social media. In: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM).

eMarketer, 2014. Smartphone Users Worldwide will Total 1.75 Billion in 2014. <http://www.emarketer.com/Article/Smartphone-Users-Worldwide-Will-Total-175-Billion-2014/1010536> (accessed 3.04.14).

Eunjoon (Cho), Seth A (Myers), Jure (Leskovec), 2011. 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, pp. 1082–1090.

Farrahi, Katayoun., Gatica-Perez, Daniel., 2011. Discovering routines from large-scale human locations using probabilistic topic models. ACM Trans. Intell. Syst. Technol. 2 (1).

Gao, Huiji, Tang, Jiliang, Liu, Huan, 2012. gSCorr: modeling geo-social correlations for new check-ins on location-based social networks. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM, pp. 1582–1586.

Griffiths, Thomas L., Steyvers, Mark., 2004. Finding scientific topics. Proc. Natl. Acad. Sci. 101 (suppl. 1), 5228–5235.

Halevy, Alon, Norvig, Peter, Pereira, Fernando, 2009. The unreasonable effectiveness of data. IEEE Intell. Syst. 24 (2), 8–12.

Hanson, Susan, Hanson, Perry, 1981. The travel-activity patterns of urban residents: dimensions and relationships to sociodemographic characteristics. Econ. Geograph. 57 (4), 332–347.

Hasan Samiul, 2013. Modeling Urban Mobility Dynamics Using Geo-location Data. Ph.D. Dissertation, Purdue University.

Hasan, Samiul, Zhan, Xianyuan, Ukkusuri, Satish V., 2013. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. ACM.

Huynh, Tâm, Fritz, Mario, Schiele, Bernt, 2008. Discovery of activity patterns using topic models. In: Proceedings of the 10th International Conference on Ubiquitous Computing. ACM, pp. 10–19.

Jin, Peter J., Cebelak, Meredith, Yang, Fan, Ran, Bin, Walton, C. Michael, Zhang, Jian, 2014. Location-based social networking data: exploration of use of doubly constrained gravity model for origin-destination estimation. In: The 93rd Annual Meeting of Transportation Research Board.

Joh, Chang-Hyeon, Arentze, Theo, Timmermans, Harry, 2001. Pattern recognition in complex activity travel patterns: comparison of Euclidean distance, signal-processing theoretical, and multidimensional sequence alignment methods. Transport. Res. Record J. Transport. Res. Board 1752 (1), 16–22.

Joh, Chang-Hyeon, Arentze, Theo, Hofman, Frank, Timmermans, Harry, 2002. Activity pattern similarity: a multidimensional sequence alignment method. Transport. Res. Part B: Methodol. 36 (5), 385–403.

Koppelman, Frank S., Pas, Eric I., 1983. Travel-activity behavior in time and space: methods for representation and analysis. School of Engineering, Duke University.

Ni, Ming, He, Qing, Gao, Jing, 2014. Using social media to predict traffic flow under special event conditions. In: The 93rd Annual Meeting of Transportation Research Board.

Noulas, Anastasios, Scellato, Salvatore, Mascolo, Cecilia, Pontil, Massimiliano, 2011. An empirical study of geographic user activity patterns in foursquare. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM).

Pas, Eric I., 1982. Analytically derived classifications of daily travel-activity behavior: description, evaluation, and interpretation. Transport. Res. Record..

Pas, Eric I., 1983. A flexible and integrated methodology for analytical classification of daily travel-activity behavior. Transport. Sci. 17 (4), 405–429.

Pas, Eric I., 1984. The effect of selected sociodemographic characteristics on daily travel-activity behavior. Environ. Plan. A 16 (5), 571–581.

Recker, Wilfred W., McNally, Michael G., Root, Gregory S., 1985. Travel/activity analysis: pattern recognition, classification and interpretation. Transport. Res. Part A: General 19 (4), 279–296.

Robert, Christian, 2001. The bayesian choice: from decision-theoretic foundations to computational implementation. In: Springer Texts in Statistics. Springer.

Sammour, George, Bellemans, Tom, Vanhoof, Koen, Janssens, Davy, Kochan, Bruno, Wets, Geert, 2012. The usefulness of the sequence alignment methods in validating rule-based activity-based forecasting models. Transportation 39 (4), 773–789.

Steyvers, Mark., 2010. Combining feature norms and text data with topic models. Acta Psychol. 133 (3), 234–243.

Steyvers, Mark, Griffiths, Tom, 2007. Probabilistic topic models. Handbook of Latent Semantic Analysis, vol. 427(7), pp. 424–440.

Wilson, Clarke., 2008. Activity patterns in space and time: calculating representative Hagerstrand trajectories. Transportation 35 (4), 485–499, http://dx.doi.org/10.1007/s11116-008-9162-z.

Yang, Fan, Jin, Peter J., Wan, Xia, Li, Rui, Ran, Bin, 2014. Dynamic origin-destination travel demand estimation using location based social networking data. In: The 93rd Annual Meeting of Transportation Research Board.