

# A Probabilistic Approach to Mining Mobile Phone Data Sequences

Katayoun Farrahi

JKU University Linz, Austria

farrahi@pervasive.jku.at

· Daniel Gatica-Perez

Idiap Research Institute, Martigny, Switzerland

EPFL, Lausanne, Switzerland

gatica@idiap.ch

the date of receipt and acceptance should be inserted later

**Abstract** We present a new approach to address the problem of large sequence mining from big data. The particular problem of interest is the effective mining of long sequences from large-scale location data to be practical for Reality Mining applications, which suffer from large amounts of noise and lack of ground truth. To address this complex data, we propose an unsupervised probabilistic topic model called the distant n-gram topic model (DNTM). The DNTM is based on Latent Dirichlet Allocation (LDA), which is extended to integrate sequential information. We define the generative process for the model, derive the inference procedure, and evaluate our model on both synthetic data and real mobile phone data. We consider two different mobile phone datasets containing natural human mobility patterns obtained by location sensing, the first considering GPS/wifi locations and the second considering cell tower connections. The DNTM discovers meaningful topics on the synthetic data as well as the two mobile phone datasets. Finally, the DNTM is compared to LDA by considering log-likelihood performance on unseen data, showing the predictive power of the model. The results show that the DNTM consistently outperforms LDA as the sequence length increases.

## 1 Introduction

As large scale mobile phone datasets on human behavior become more readily available, the need for effective methods and mathematical models for analysis becomes crucial. Research in Reality Mining [7, 10] has led to the need for the development of models that discover patterns over long and potentially varying durations. We address the problem of modeling long duration activity sequences for large-scale human routine discovery from cellphone sensor data. Our

objective is to handle sequences corresponding to human routines based on principled procedures, and to apply them to human location data.

There are several difficulties to modeling human activities, including various types of uncertainty, lack of ground truth, complexity due to the size of the data, and diversity of phone users. One fundamental issue motivating this work is that we often do not know (or cannot pre-specify) the basic units of time for the activities in question. We do know that human routines have multiple timescales (hourly, daily, etc.), however the effective modeling of multiple unknown time durations is an open problem. Secondly, the problem of mining location sequences quickly results in an exponential number of possibilities, particularly when considering the wide range of locations visited by people and the order in which the locations occur. The focus of our model is to address the issue of modeling long sequences (such as those occurring in mobility patterns) by proposing a novel approach based on latent topics in order to avoid parameter dimension explosion.

We focus on probabilistic topic models as the basic tool for routine analysis for several reasons. Topic models are, first and foremost, unsupervised in nature. Their probabilistic generative nature make them attractive over discriminative approaches since we are interested in mining the structure of the data. Topic models are also intuitive and provide opportunity for extensions with approximate methods for inference. They can handle uncertainty due to the exchangeability of the bag of words property and process large amounts of data [25]. They can also be extended in various ways to integrate multiple data types [10].

The contributions of this paper are as follows: (1) we propose the distant n-gram topic model (DNTM) for sequence modeling; (2) we derive the inference process using Markov Chain Monte Carlo (MCMC) sampling [21]; (3) we generate a dataset of synthetic sequences and ap-

ply the DNTM to test the model under a controlled setting; (4) we apply the DNTM to two real large-scale mobile phone location datasets. The model discovers user location routines over several hour time intervals, corresponding to sequences, and these results are illustrated by differing means; (5) we also perform a comparative analysis with Latent Dirichlet Allocation (LDA) [4], showing that the DNTM performs better in predicting unseen data based on log-likelihood values. This paper is an extended version of the work originally presented at [11].

This paper is structured as follows. We begin by presenting the most related work in Section 2. We introduce the Distant N-Gram Topic Model (DNTM) in Section 3, defining the graphical model, the generative procedure, and the inference and parameter estimation details. We then evaluate the DNTM on a synthetic dataset in Section 4 followed by two real mobile phone datasets in Section 5. We conclude with a discussion followed by the conclusion and future works.

## 2 Related Work

This section discusses related work in mobility modeling methods for location data from cell phones and on probabilistic topic models.

### 2.1 Mobility Patterns from Phone Data

There have been many recent works considering large-scale mobile phone calling occurrences to obtain location data from cell tower connections. Such datasets are available to mobile phone operators and contain sparse location information over a large set of users. We consider this data to be sparse since location is only available when a phone call takes place, otherwise the location is unknown. Based on this data, several problems relating to activity modeling have been addressed.

Phithakkitnukoon et al. [26] identify daily human activity patterns of eating, shopping, entertainment, and recreation from location estimates at the beginning and end of calls, messages, and internet connections over a data collection of one million users over a few month period. Candia et al. [5] propose an approach to discover what they refer to as spatio-temporal anomalies, which are anomalous events in the mean collective behavior of individuals obtained by resolving phone call records in time and space. Gonzalez et al. [12] find that human trajectories show a high degree of temporal and spatial regularity by considering a data collection of 100,000 users over a 6 month period. They find that each individual can be characterized by a time-independent travel distance and has a significant probability to return to a few highly frequented locations. The most closely related work to ours in this category is by Gornepur [13], in

which a probabilistic approach for mining common routes from cell tower IDs is presented. This paper extends the work by Becker et al. [3] by addressing scalability. The approach considers two steps, the first is the locality-sensitive hashing of the cell ID sequences, which disregards the order of cell ID occurrences. The second step is graph clustering resulting in groups of cell ID sequences. The work is evaluated with GPS traces collected by the author. The main advantages of the approach are its scalability and the resulting anonymisation of personal trajectory information. The drawbacks are that the work was evaluated on a small dataset with a small set of routes and base station densities. Further, the disregard of the cell ordering information simplifies the method, particularly since time of day information is not considered. In our work, however, the goal is not to obtain route information from cell tower ID sequences, but to mine dominantly occurring sequences of locations.

The problem of mobility modeling directly using mobile location sensor data has also been studied previously. Previous work by Zheng et al. has been done to mine locations of interest and top travel sequences in a geospatial region [32]. The approach additionally infers the most experienced users in a geo-related community using GPS trajectories. The algorithm links users and locations, where users point to many locations and locations are pointed to by many users. These weighted links are used to mine the locations of interest and determine the top travel sequences and the main application of the work is location recommendation. Hightower et al. [15] use WiFi and GSM radio fingerprints collected by personal mobile devices to automatically learn places and then to detect when users will return to those places. Their algorithm is called BeaconPrint and is compared to three similar previous strategies [1, 18, 22]. They conclude that BeaconPrint is 90% accurate in learning and recognizing the places people visit. An unsupervised approach based on particle filters has been developed by Patterson et al. [24] to simultaneously learn a unified model of transportation modes as well as most likely routes. The data considered is taken from a GPS sensor stream collected by the authors over a period of three months. Yavas et al. [30] present a data mining algorithm for the prediction of user movements. The algorithm proposed is based on mining the mobility patterns of users, forming mobility rules from these patterns, and finally predicting users' next movements.

Our overall goal of mining location sequences and the latent topic modeling approach in this paper differ from these previous works, which also considered location sensor data for activity modeling.

### 2.2 Topic Models

Probabilistic topic models were initially developed to analyze large collections of text documents [4, 16]. They have

been used more recently for other sources of data such as location [10] and physical proximity [2,6]. Here we consider their application to large scale mobile phone data.

Previously, we used existing topic models (Probabilistic Latent Semantic Analysis, LDA, and the Author Topic Model) [8,10] for human activity discovery and prediction using cell tower and a small collection of GPS data. This paper extends on this initial work by defining in detail a new model to address the limitation of long duration activity discovery with topic models.

Bao et al. [2] address a similar problem (modeling user mobile contexts) with unsupervised models, namely an extension of LDA. However, the focus of [2] is on incorporating dependencies among context, features, and external conditions into the model. Huynh et al. [17] use LDA for activity recognition, but considering wearable sensors and considering fine-grained daily activities such as washing hands. Do and Gatica-Perez [6] introduce a topic model for group discovery from Bluetooth interaction data. They develop an unsupervised topic model based on LDA which discovers dominantly co-occurring group interaction patterns over time. Recently, Zheng and Li [31] proposed an unsupervised approach to mine location-driven activities to enable activity discovery from celltowers. Time is modeled explicitly, and the model can be used for location prediction. The model can compare users' activities as well. However, sequential information is discarded by the model (due to the bag of words), and the focus is prediction and user comparison. None of these previous works focuses on the issue we address in this paper, namely to model sequence information using topic models in a manner that can handle long sequences, which is necessary for human activities.

Topic models have previously been used for  $n$ -gram discovery in the context of text and speech. The bigram topic model [28], the LDA collocation model [29], and the topical  $n$ -gram model [29] are all extensions of LDA to tackle this problem. The topical  $n$ -gram model is an extension to the LDA collocation model, and is more general than the bigram model. This approach was developed to be applied to text modeling, and retains counts of bigram occurrences, and thus could not easily be extended for large  $n$  (i.e.  $n > 3$ ) due to parameter dimension explosion. The multi-level topic model is another extension of LDA for  $n$ -gram discovery [9], cascading a series of LDA blocks for varying length sequence discovery. The problem of activity discovery from mobile phone data requires  $n$ -gram models capable of handling long sequences; we approach this issue by modeling a simplified dependency between labels (or words) within a sequence and adding a dependency to topics; we find that this technique is promising for location sequence discovery.

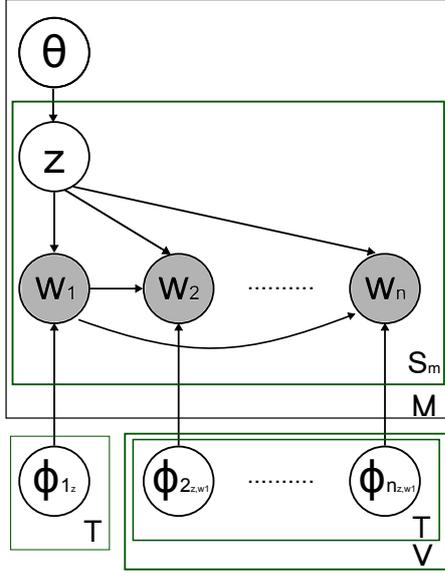
### 3 Distant N-Gram Topic Model

#### 3.1 Topic Models Basics

Latent Dirichlet Allocation (LDA) [4] is a generative model in which each document is modeled as a multinomial distribution of topics and each topic is modeled as a multinomial distribution of words. By defining a Dirichlet prior on the document/topic ( $\Theta$ ) and word/topic ( $\Phi$ ) distributions, LDA provides a statistical foundation and a proper generative process. The main objective of the inference process is to determine the probability of each word given each topic, resulting in the matrix of parameters  $\Phi$ , as well as to determine the probability of each topic given each document, resulting in  $\Theta$ . Formally, the entity termed *word* is the basic unit of discrete data defined to be an item from a vocabulary. In the context of this paper, a word, later referred to as a label  $\mathbf{w}$ , is analogous to a person's location. A *document* is a collection of words also referred to as a bag of words. In our case, a document is a day in the life of an individual. A *corpus* is a collection of  $M$  documents. In this paper, a corpus corresponds to the collection of sensor data to be mined. In the context of text, a *topic* can be thought of as a 'theme', whereas in our analogy, a topic can be interpreted as a human location routine.

#### 3.2 DNTM Overview

We introduce a new probabilistic generative model for sequence representation. The model is built on LDA, with the extension of generating sequences instead of single words as LDA does. The limiting criteria is to avoid parameter dimension explosion. We define a sequence to be a series of  $N$  consecutive labels or words. We represent a sequence as follows:  $\mathbf{q} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_N)$ , where  $\mathbf{w}$  denotes a label. In the context of this paper, a label  $\mathbf{w}$  corresponds to a user's location obtained from a mobile phone sensor, though in general a label can correspond to any given feature in a series. The sequence  $\mathbf{q}$  is then a sequence of locations occurring over an interval of time. The interval of time is defined by the duration over which each label occurs times the number of elements  $N$  in the sequence. The distant  $n$ -gram topic model (DNTM) defines a generative process for a corpus of sequences. The maximum length of the sequence  $N$  is predefined. In existing  $n$ -gram models [29], a label in a sequence is assumed to be conditionally dependent on all previous labels in the sequence, thus making large sequences (longer than 3 labels) infeasible to manage due to an exponential number of dependencies as the sequence length grows. In contrast here, we integrate latent topics and assume a label in the sequence to be conditionally dependent only on the first element, the distance to this label, and the correspond-



**Fig. 1** Graphical model of the Distant N-Gram Topic Model (DNTM). A sequence  $\mathbf{q}$  is defined to be  $N$  consecutive locations  $\mathbf{q} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_N)$ . Latent topics,  $z$ , are inferred by the model and can be interpreted as the different routines found to dominate the sensor data. There are  $M$  days (or documents) in the dataset.  $\theta$  is a distribution of days given the routines, and  $\Phi_j$  is a distribution of location sequences given routines.

ing topic, removing the dependency on all other labels, and thus removing the exponential parameter growth rate.

The underlying concept and the novelty of our method is to obtain a distribution of topics given the first element in a sequence, represented by  $\Phi_{1z}$ . Then for each position  $j$  in the sequence, where  $j > 1$ , the distribution of topics given the  $j^{\text{th}}$  position in the sequence is obtained, depending on both the first element and the topic, represented by  $\Phi_{jz,w_1}$ . With this logic, our parameter size grows linearly with the sequence length  $N$ . Note that our approach for label dependency on  $w_1$  is the simplest case for which a label is always present. More advanced methods, including determining the number of previous labels for dependency are the subject of future work. We apply this model to location data to discover activities over large durations considering intervals of up to several hours. Next we define the generative process and introduce the learning and inference procedure. More derivation details can be seen in the Appendix, and the full derivation can be found in [8] where our model was referred to with a slightly different acronym.

### 3.3 The Probabilistic Model

The graphical model for our distant n-gram topic model is illustrated in Figure 1. We use a probabilistic approach where

**Table 1** Symbol description

$N$	The length of the sequence
$\mathbf{q}$	A sequence of $N$ consecutive labels ( $\mathbf{w}_1, \dots, \mathbf{w}_N$ )
$m$	An instance of a document (a day here)
$S_m$	The total number of sequences $\mathbf{q}$ in document $m$
$M$	The number of documents in the corpus
$T$	The number of latent topics
$z$	A latent topic (a location routine here)
$V$	The vocabulary size
$\theta$	The distribution of topics given documents
$\Phi$	The distribution of sequences given topics, where $\Phi = \{\Phi_{1z}, \Phi_{2z,w_1}, \dots, \Phi_{nz,w_1}\}$
$\Phi_{1z}$	The distribution of $w_1$ given topics
$\Phi_{jz,w_1}$	The distribution of $w_j$ given $w_1$ and topics

observations are represented by random variables, highlighted in gray. The latent variable  $z$  corresponds to a topic of activity sequences. The model parameters are defined in Table 1.

The generative process is defined as follows:

#### 1. Initialization:

(1) For each document  $m$  in the corpus draw a distribution over topics  $\theta_m \sim \text{Dirichlet}(\alpha)$ .

(2) For each document  $m$  in the corpus:

(2.1) For each sequence  $\mathbf{q}$  in document  $m$ :

(2.1.1) Draw a distribution over labels

$\Phi_{1z} \sim \text{Dirichlet}(\beta_1)$  for the first element in the sequence.

(2.1.2) For each consecutive label  $w_j$  in the sequence:

Draw a distribution over labels  $\Phi_{jz,w_1} \sim \text{Dirichlet}(\beta_j)$ . Here  $\Phi_{jz,w_1}$  captures the dependency with  $z$ ,  $w_1$ , as well as the distance from the first label. Note the sequence length is defined by the user and is fixed.

#### 2. Sequence generation procedure.

(1) For each document  $m$  in the corpus:

(1.1) For each sequence  $\mathbf{q}$  of the  $S_m$  sequences in document  $m$ :

(1.1.1) Draw a topic  $z|m \sim \text{Multinomial}(\theta_m)$ .

(1.1.2) Draw the first label in the sequence  $w_1|z \sim \text{Multinomial}(\Phi_{1z})$ .

(1.1.3) For  $j = 2$  to  $N$ :

Draw the  $j$ -th label in the sequence  $w_j|w_1, z \sim \text{Multinomial}(\Phi_{jz,w_1})$  for  $1 < j \leq N$ .

In summary, in the generative process for each sequence, the model first picks the topic  $z$  of the sequence and then generates all the labels in the sequence. The first label in the sequence is generated according to a multinomial distribution  $\Phi_{1z}$ , specific to the topic  $z$ . The remaining labels in the sequence,  $w_j$  for  $1 < j \leq N$ , are generated according to a multinomial  $\Phi_{jz,w_1}$  specific to the current label position  $j$ , the topic  $z$  as well as the first label of the sequence  $w_1$ . Note  $j$  is the  $j$ -th label in the sequence, but it can also be viewed as the distance between label  $j$  and 1.

$$\begin{aligned}
p(\mathbf{z}, \mathbf{q} | \alpha, \beta) &= p(\mathbf{z} | \alpha) p(\mathbf{w}_1 | \mathbf{z}, \beta_1) \prod_{j=2}^n p(\mathbf{w}_j | \mathbf{z}, \mathbf{w}_1, \beta_j) \quad (1) \\
&= \int_{\Theta} p(\mathbf{z} | \Theta) p(\Theta | \alpha) d\Theta \cdot \int_{\Phi_1} p(\mathbf{w}_1 | \mathbf{z}, \Phi_1) p(\Phi_1 | \beta_1) d\Phi_1 \cdot \\
&\quad \prod_{j=2}^n \int_{\Phi_j} p(\mathbf{w}_j | \mathbf{w}_1, \mathbf{z}, \Phi_j) p(\Phi_j | \beta_j) d\Phi_j \quad (2) \\
&= \prod_{m=1}^M \frac{B(n_m + \alpha)}{B(\alpha)} \cdot \prod_{k=1}^T \left( \frac{B(n_k + \beta_1)}{B(\beta_1)} \right) \cdot \prod_{j=2}^n \frac{B(n_{k_j} + \beta_j)}{B(\beta_j)} \quad (3)
\end{aligned}$$

We define the following notation;  $n_m^k$  is the number of occurrences of topic  $k$  in document  $m$ ;  $n_m = \{n_m^k\}_{k=1}^T$ ;  $n_k^{w_1}$  is the number of occurrences of label  $w_1$  in topic  $k$ ,  $n_k = \{n_k^t\}_{t=1}^V$ ; finally  $n_{k_j}^{(w_1, w_2)_j}$  is the number of occurrences of label  $w_2$  occurring  $j$  labels after  $w_1$  in topic  $k$  and  $n_{k_j} = \{n_{k_j}^{(w_1, w_2)_j}\}_{w_1=1, w_2=1}^{V, V}$ .

We assume a Dirichlet prior distribution for  $\Theta$  and  $\Phi = \{\Phi_{1, \mathbf{z}}, \Phi_{2, \mathbf{z}, w_1}, \dots, \Phi_{n, \mathbf{z}, w_1}\}$  with hyperparameters  $\alpha$  and  $\beta = \{\beta_1, \beta_2, \dots, \beta_n\}$ , respectively. We assume symmetric Dirichlet distributions with scalar parameters  $\alpha$  and  $\beta$  such that  $\alpha = \sum_{k=1}^T \frac{\alpha_k}{T}$ ,  $\beta_1 = \sum_{v=1}^V \frac{\beta_{1,v}}{V}$ , and  $\beta_j = \sum_{w_1=1}^V \sum_{w_2=1}^V \frac{\beta_{(w_1, w_2)_j}}{V^2}$  for  $1 < j \leq N$ . Note the parameters  $\alpha_k$ ,  $\beta_{1,v}$ , and  $\beta_{(w_1, w_2)_j}$  are the components of the hyperparameters  $\alpha$ ,  $\beta_1$ , and  $\beta_j$ , respectively in the case of non-symmetric Dirichlet distributions. The joint probability of observations and latent topics can be obtained by marginalizing over the hidden parameters  $\Theta$  and  $\Phi$ . These relations are then used for inference and parameter estimation in Equations (1)-(3), where  $p(\mathbf{z} | \alpha)$ ,  $p(\mathbf{w}_1 | \mathbf{z}, \beta_1)$ , and  $p(\mathbf{w}_j | \mathbf{w}_1, \beta_j)$  resulting in the following. Note, derivation details can be found in the appendix and in [8].

$$p(\mathbf{z} | \alpha) = \prod_{m=1}^M \frac{B(n_m + \alpha)}{B(\alpha)} \quad (4)$$

where  $n_m = \{n_m^k\}_{k=1}^T$

$$p(\mathbf{w}_1 | \mathbf{z}, \beta_1) = \prod_{k=1}^T \frac{B(n_k + \beta_1)}{B(\beta_1)} \quad (5)$$

where  $n_k = \{n_k^t\}_{t=1}^V$

and for  $1 < j \leq n$

$$p(\mathbf{w}_j | \mathbf{w}_1, \mathbf{z}, \beta_j) = \prod_{k=1}^T \frac{B(n_{k_j} + \beta_j)}{B(\beta_j)} \quad (6)$$

where  $n_{k_j} = \{n_{k_j}^{(t_1, t_2)_j}\}_{t_1=1, t_2=1}^{V, V}$

### 3.4 Inference and Parameter Estimation

Like LDA, the optimal estimation of model parameters is intractable. The model parameters are derived based on the MCMC approach of Gibbs sampling [14]. The model parameters can then be estimated by solving the following relationship.

$$p(z_i = k | \mathbf{z}_{-i}, \mathbf{q}, \alpha, \beta) \propto (n_{m, -i}^k + \alpha) \cdot \frac{n_{k, -i}^t + \beta_1}{\sum_{t=1}^V n_{k, -i}^t + \beta_1} \cdot \prod_{j=2}^n \frac{n_{k, -i}^{(t_1, t_2)_j} + \beta_j}{\sum_{t_1=1}^V \sum_{t_2=1}^V n_{k, -i}^{(t_1, t_2)_j} + \beta_j} \quad (7)$$

where  $n_x^{(y)} = n_{x, -i}^{(y)} + 1$  if  $x = x_i$  and  $y = y_i$  and  $n_x^{(y)} = n_{x, -i}^{(y)}$  in other cases.

The model parameters can then be estimated by sampling the dataset using the following relations:

$$\theta_m^k = \frac{n_m^k + \alpha}{\sum_{k=1}^T (n_m^k + \alpha)} \quad (8)$$

$$\phi_{1, k}^t = \frac{n_k^t + \beta_1}{\sum_{t=1}^V (n_k^t + \beta_1)} \quad (9)$$

$$\phi_{j, k}^{(t_1, t_2)_j} = \frac{n_k^{(t_1, t_2)_j} + \beta_j}{\sum_{t_1=1}^V \sum_{t_2=1}^V (n_k^{(t_1, t_2)_j} + \beta_j)} \quad (10)$$

where  $n_k = \{n_k^t\}_{t=1}^V$  and  $n_{k_j} = \{n_{k_j}^{(t_1, t_2)_j}\}_{t_1=1, t_2=1}^{t_1=V, t_2=V}$ .

## 4 Experiments on Synthetic Data

First we consider synthetic data to demonstrate the strength of the DNTM. We consider a vocabulary of 10 possible location labels  $w_i$  thus  $V = 10$ . We first create 5 topics each represented as a sequence of 6 location labels inspired by the synthetic topics developed in [27]. We create one document of 2000 random sequences assuming equi-probable topics following the generative process of Section 3.3. The five topics are shown in Figure 3, where each topic contains a sequence of 6 location labels (x-axis). Note, in topic 4 there is an equal probability of generating a sequence with labels 1-9 in position 3 but not label 10.

The topics learned by the DNTM are shown in Figure 4 for  $N = 6$ ,  $T = 5$ , and  $\alpha = 0.1$ ,  $\beta = 0.1$  (we assume all  $\beta_i$  are equal to  $\beta$ ). We plot the most probable location label for each position in the sequence given the topic, (i.e.  $p(w_j | w_1, k)$  for position  $j$  and topic  $k$ ). The x-axis corresponds to the sequence position and the y-axis to the possible location labels. We reorder the topics learned by the model to correspond to the topics in Figure 3. We plot the 10 most probable sequences discovered by the model for topic 1 (corresponding to Topic 4 Figure 3) in order to illustrate

```

// GOAL: Given a training corpus,  $\alpha, \beta, T$  and  $n$ , estimate the parameters  $n_m^k, n_k^t$ , and  $n_k^{(t_1, t_2)j}$  for
//  $j = 2$  to  $n$  from which we can determine the model parameters  $\theta_m^k, \phi_{1,k}^t$ , and  $\phi_{j,k}^{(t_1, t_2)j}$ .

// Initialization
1) Initialize the count parameters,  $n_m^k = 0, n_k^t = 0, n_k^{(t_1, t_2)j} = 0$  for  $j = 2$  to  $n$ .
2) Iterate over each sequence  $\mathbf{q}$  in the corpus:
    3) Sample a topic  $k$  from  $k \sim \text{Mult}(\frac{1}{T})$ .
    4) Update the count parameters  $n_m^k, n_k^t, n_k^{(t_1, t_2)j}$  as follows  $n_m^k = n_m^k + 1, n_k^t = n_k^t + 1,$ 
        $n_k^{(t_1, t_2)j} = n_k^{(t_1, t_2)j} + 1$  for  $j = 2$  to  $n$ .

// Run the chain
5) Iterate over a large number of iterations (e.g. 1000):
    6) Iterate over each sequence:
        7) Decrement the current sequence and sequence elements' topic assignments as follows
            $n_m^k = n_m^k - 1, n_k^t = n_k^t - 1, n_k^{(t_1, t_2)j} = n_k^{(t_1, t_2)j} - 1$  for  $j = 2$  to  $n$ .
        8) Sample a topic  $k$  for the sequence from  $p(z = k | \mathbf{z}_{-i}, \mathbf{w}) \propto (n_{m,-i}^k + \alpha) \cdot$ 
            $\frac{n_{k,-i}^t + \beta_1}{\sum_{t=1}^T n_{k,-i}^t + \beta_1} \cdot \prod_{j=2}^n \frac{n_{k,-i}^{(t_1, t_2)j} + \beta_j}{\sum_{t_1=1}^T \sum_{t_2=1}^T n_{k,-i}^{(t_1, t_2)j} + \beta_j}$ .
        9) Increment the new topic assignments as follows  $n_m^k = n_m^k + 1, n_k^t = n_k^t + 1,$ 
            $n_k^{(t_1, t_2)j} = n_k^{(t_1, t_2)j} + 1$  for  $j = 2$  to  $n$ .

// Compute model parameters
10) Estimate the unknown parameters as follows
 $\theta_m^k = \frac{n_m^k + \alpha}{\sum_{k=1}^T (n_m^k + \alpha)}, \phi_{1,k}^t = \frac{n_k^t + \beta_1}{\sum_{t=1}^T (n_k^t + \beta_1)}$ , and
 $\phi_{j,k}^{(t_1, t_2)j} = \frac{n_k^{(t_1, t_2)j} + \beta_j}{\sum_{t_1=1}^T \sum_{t_2=1}^T (n_k^{(t_1, t_2)j} + \beta_j)}$ , for  $j = 2$  to  $n$ .

```

Fig. 2 Gibbs Sampling Algorithm for the Pairwise-Distance Topic Model.

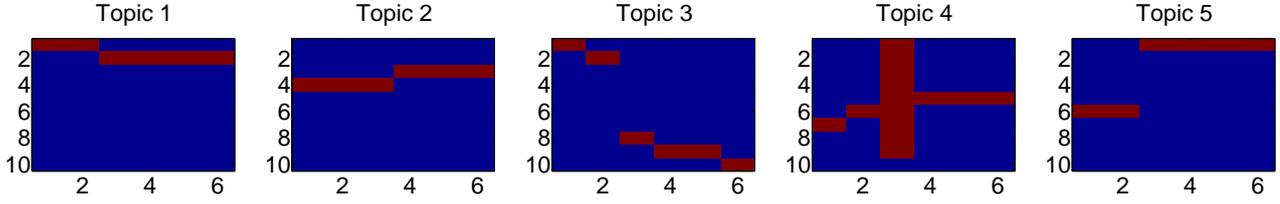


Fig. 3 Synthetic sequences to test the distant n-gram topic model. Each topic contains one sequence of length 6 (x-axis). There are 10 possible location labels (y-axis). Note position 3 in topic 4.

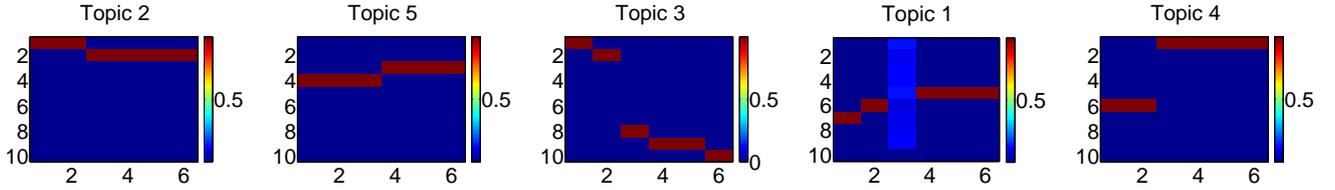


Fig. 4 DNTM results for  $N = 6, T = 5, \alpha = 0.1, \beta = 0.1$ . All of the sequences discovered by the DNTM correspond to the correct synthesized topics presented in Figure 3. The colorbar displays the probability of the sequence elements given the topics. Note the correct discovery of locations 1-9 but not 10 in topic 1 position 3.

the model correctly learned the 9 possible locations for position 3.

Next, we consider a more complex synthetic dataset of 6 topics consisting of multi-length sequences  $N = 6$  and  $N = 9$ . These topics are shown in Figure 6, where topic 4 (d) and topic 5 (e) contain the sequences of length 9 and (a)-(c) and (f) are length 6. Again, one document is generated by

randomly sampling the topics (a)-(f) with equal probability following the generative process of Section 3.3. We refer to this test set as the multi-length synthetic data.

The results for the DNTM on the multi-length synthetic data are shown in Figure 6. The DNTM is run with  $N = 12, T = 10$ , and  $\alpha = 0.1, \beta = 0.1$ . In Figure 6 we show the single most probable sequence discovered by the model

for select topics, corresponding to topics in Figure 5. We consider  $N = 12$  in order to capture all of the sequences. When  $N = 9$ , all of the sequences of length 6 are discovered, however the sequences of length 9 are cut up between topics. By setting  $N = 12$  the sequences of length 9 are not cut up between topics and occur within the interval of length 12 often enough in order for the model to capture the co-occurrences. Since Figure 6 displays the sequences of length 12 discovered, segments of other sequences are also discovered which often co-occurred with the sequences; all of the input sequences are correctly discovered by the DNTM. Note the colorbar displays the probability of the location element given the topic. Considering the sequence 112222 (Figure 5 (a)), it appears in position 4 to 10 (Figure 6 (a)). Similarly sequence (b) 444333 appears in position 1 to 6, and so on. Note topic 10 (Figure 5 (d)) contains a small probability of possible locations for position 6 though we just plot the single most probable sequence.

## 5 Mobile Phone Location Data

The DNTM could be potentially applied to any type of data with discrete valued labels in a sequence, for example text. We are interested in mobile location data over time. As stated in Section 2, we make an analogy with LDA where a document is an interval of time in a person’s daily life. Here we always consider a document to be a day in the life of a user. A label  $\mathbf{w} = (t, l)$  is composed of a location  $l \in L$ , where  $L$  is the discrete set of possible locations which occurred over a 30 minute interval and a time coordinate of the day  $t \in Z = \{1, 2, 3, \dots, tt\}$ . We consider two different datasets for experiments. The representations for each are detailed below.

### 5.1 Nokia Smartphone Data

We use real life data from 25 users using a Nokia N95 smartphone from 2009.10.01 to 2010.07.01 corresponding to a nine-month period of the Lausanne Data Collection Campaign [19]. The phone has an application that collects location data on a quasi-continuous basis using a combination of GPS and WiFi sensing, along with a method to reduce battery consumption. Place extraction was performed using the algorithm proposed in [23], that reported good performance on similar data. The place extraction algorithm is described in more detail in the next subsection (Section 5.1.1). In Section 6.1, we create  $\mathbf{w}$  where  $tt = 8$ , (i.e., the day is divided into 8 equivalent time intervals),  $L = \{l_0, l_1, l_2, \dots, l_{MAX}\}$ , where  $MAX$  is the number of detected places determined by [23], and  $l_i$  is the user-specific index of the place. In Section 6.2, we study a second case in which we disregard  $tt$ .

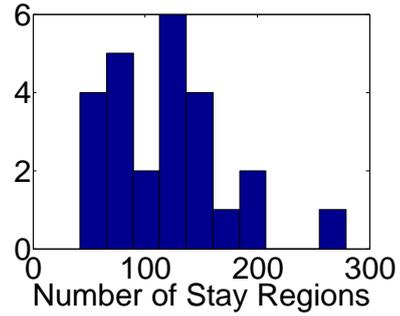


Fig. 7 Histogram of the number of stay regions per user.

If  $l_i = 0$ , there is no detected place, either due to no location being sensed, or due to the user moving or not staying at the location for very long. All places  $l_i > 0$  are indexed according to their frequency of occurrence. Note that each user has a differing set of places and for this data collection, topics are discovered on an individual basis. We show the histogram of  $l_{MAX}$  over the 25 users in Figure 7. One user has a much larger number of stay regions than the majority. The average number of stay regions for this group of users is 117.5.

#### 5.1.1 Place Extraction Algorithm

Place extraction was performed on the location data using the algorithm in [23] in order to obtain a manageable number of regions of interest frequented by users from the large number of location points sensed. The algorithm has two levels of clustering. The location coordinates are first clustered into stay points, where stay points are clusters of coordinates from the same day, representing geographic regions in which a user stayed for a while. Stay points are then clustered into stay regions, where stay regions are places of interest from several days of data with the same semantic meaning. The purpose of this step is to reduce the large number of locations sensed for each user into a more manageable set of regions for which the user stayed in for a minimum duration of time and to disregard the regions which were not frequently visited in order to maintain a reasonable vocabulary size for the model.

In Figure 8 we plot the stay regions discovered over one user’s data. White intervals indicate that no place was observed during that time interval. This user had 101 unique stay regions found by the place extraction algorithm. In Figure 9 we show 2 of the same user’s stay regions in geographic terms that correspond to public places. We only display the satellite view for anonymity reasons.

### 5.2 MIT Reality Mining (RM) Data

The MIT RM data collected by Eagle and Pentland [7] contains the data of 97 users over 16 months in 2004-2005. This

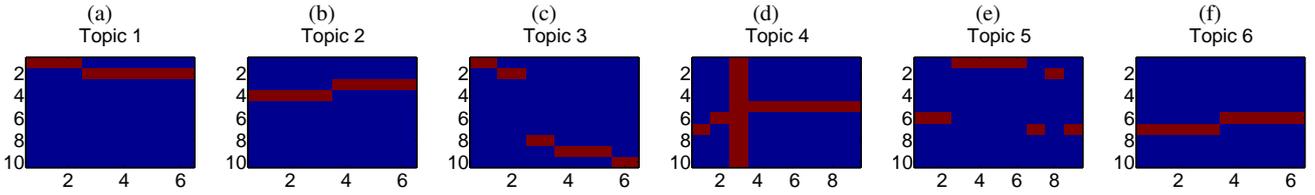


Fig. 5 Synthetic sequences of length 6 and 9 for testing the DNTM.

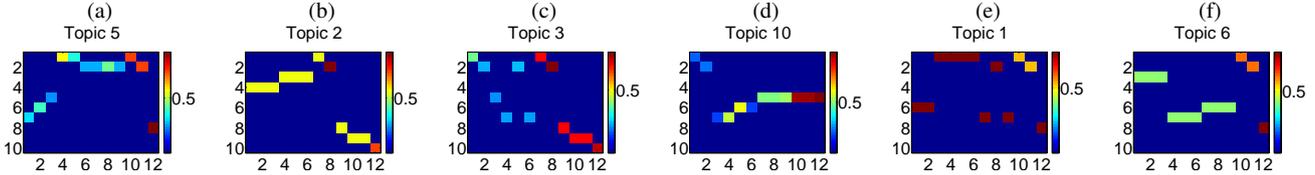


Fig. 6 DNTM results for  $N = 12$ ,  $T = 10$ ,  $\alpha = 0.1$ ,  $\beta = \beta_2 = 0.1$ . We plot the single most probable sequence output per topic. The colorbar indicates the location label probability for the topic.

data contains no detailed location information, but we define four possible location categories for a user collected via cell tower connections. The towers are labeled as 'home', 'work', 'out', or 'no reception', making the labels consistent over all the users. This corresponds to  $L = \{H, W, O, N\}$ . For this we set  $tt = 48$ .

## 6 Experiments and Results

We present the DNTM results on two real mobile phone data collections. First on the Nokia Smartphone data considering a scenario with a time coordinate of the day  $t$  in the label definition  $w$ . Then, we consider the modified scenario without a time coordinate in the vocabulary. Finally, results are presented on the MIT RM dataset.

### 6.1 Nokia Smartphone Data

For experiments with the smartphone data, we remove days that do not have at least one place detected. The results shown here are for  $T = 25$ ,  $\beta_j = 0.1$ ,  $1 \leq j \leq N$  and  $\alpha = 0.1$  selected heuristically. We consider  $N = 12$  corresponding to six-hour sequences for the topics displayed here. Note that a range of values of  $T$  give similar results, the difference being that when  $T$  is small, the overall most occurring topics are discovered, and when  $T$  is larger, more specific items are found. The constraint on the hyperparameters  $\beta_j$  and  $\alpha$  are that they be smaller than the order of label/topic and document/topic counts.

Several of the topics discovered by the DNTM for the smartphone data displayed in Figure 8 are shown in Figures 10 and 11. The first parameter the model returns is  $\Theta$ , containing a probability distribution of each day in the corpus for each topic. We rank these probabilities for each topic and visualize the 10 most probable days, illustrating which

days in the data had the highest probability of the location sequences for the given topic. In Figure 10, the three figures illustrate the 10 most probable days (i.e.,  $\max(\theta_m^k)$  for a given topic  $k$ ). The x-axis corresponds to the time of day, the y-axis corresponds to days, and each unique colour corresponds to a unique place. We can see that sequences of places occurring over particular intervals of the day are discovered by the model. For example, topic 8 for user 1 corresponds to place 1 (home in magenta) occurring over most of the day.

Using data from a different user, in Figure 11 (a) we show topic 19 discovered for user 2. We also visualize the coordinates of the place as displayed below the topic. The circle indicates the location of place 1 on a satellite map view. In Figure 11 (b) we show topic 2 for user 2. Below the topic we display the mobility traces for the day 2010.02.07, which is one of the 10 most probable days for topic 2. On the satellite view, each colour corresponds to a unique location, coordinated with the colour scheme of the topic displayed.

### 6.2 Nokia Smartphone Data: Modified Scenario

We now consider a slightly different input feature format considering the Nokia data collection. A label  $l$  is simply a location occurring over a 10-minute period without time information. Previously in Section 6.1, a stay region was considered every 30 minutes and  $tt = 8$ . The input sequences are the non-overlapping location stay regions in sequence of length  $N$  for a given user. We illustrate in Figures 13-16 that the DNTM successfully discovers location routines of large sequence lengths  $N$ .

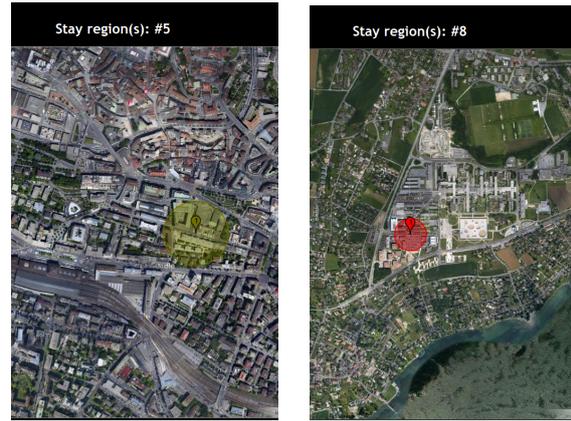
First we run the DNTM with  $T = 10$ ,  $N = 6$  for each of the 25 users and plot the probability distribution over the topics in Figure 12. Since each location label corresponds to a 10 minute interval and  $N = 6$ , we are modeling one hour sequences here. We can see that most of the probability mass



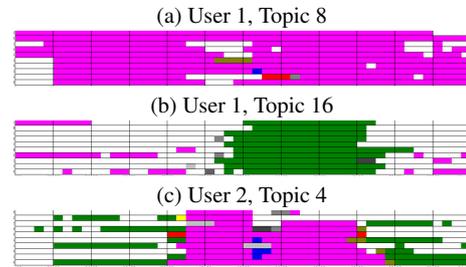
**Fig. 8** One user's data after place extraction. Each row (y-axis) corresponds to a day quantized into intervals of 10 minutes (x-axis). There are 101 unique places (stay regions) found by the place extraction algorithm. Each place is numbered according to the frequency of occurrence and assigned a unique colour. Here, pink corresponds to region 1 (home) and green corresponds to region 2 (work). Note that the regions extracted are specific to a particular user for the Nokia dataset.

is over a few topics indicating  $T$  can be smaller, however, there is no harm in setting  $T$  larger.

We plot several of the most probable topics for users by displaying the most probable days given topics,  $\Theta$  (Figure 13) and the most probable sequence given topics,  $\Phi$  (Figure 14). In Figure 13, the five most probable days are plot for each topic, where the y-axis corresponds to days, the x-axis to the time of day, and the colorbar to the locations. In Figure 14, the single most probable sequence is plot for given topics, where the y-axis corresponds to the locations, and the colorbar represents the probability of the location (or label) given the sequence position (x-axis), the first location label and the topic. The probability of each sequence component (indicated by the colorbar) differs and is an indication of the



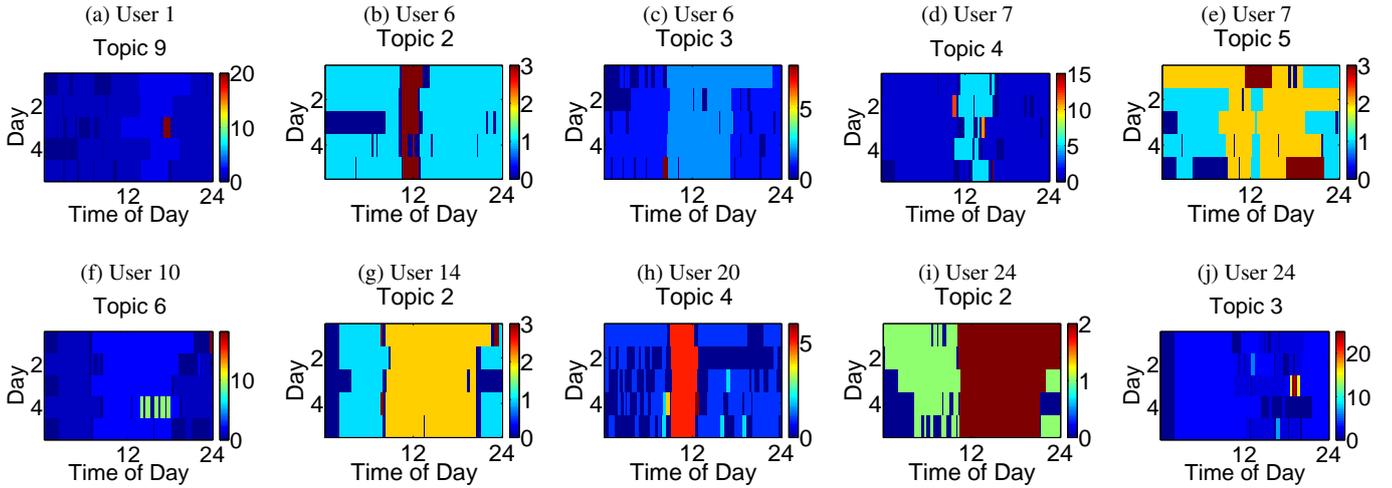
**Fig. 9** Satellite view of 2 places extracted for a user. Each color represents a given user's visited place and is used consistently across the results for the given user.



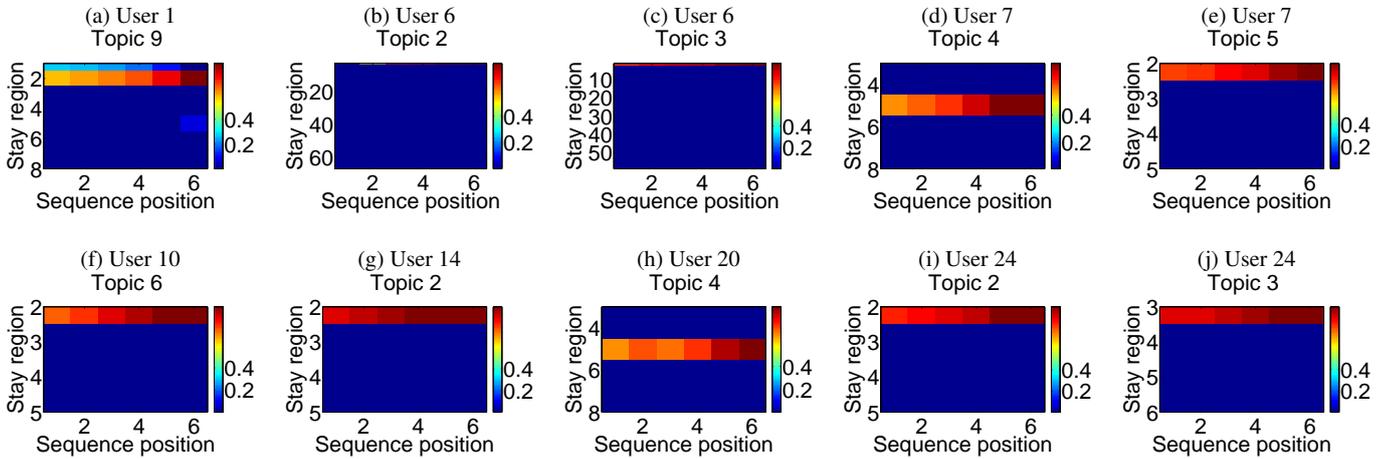
**Fig. 10** Selected topics discovered for  $N = 12$  (a,b) user 1, (c) user 2. The x-axis is the time of day, the y-axis are the 10 most probable days for the topic ranked from top to bottom (output as  $\Theta$  by the DNTM). Each unique colour represents a unique place. Our model discovers sequences of locations which dominantly occur in a user's mobility patterns. For example, topic 8 for user 1 corresponds to being at home (pink) throughout the day. Topic 16 for user 1 corresponds to being at work (green) for several hours in the afternoon.

amount of 'noise' or variation in this label occurring at this position given the topic. A wide range of mobility routines are apparent, particularly by viewing the most probable days given topics (Figure 13) for all users, where co-occurring sequential patterns of stay regions are found by the model.

Next we consider a longer sequence length,  $N = 18$  considering 3 hour sequences, with the same model parameters and again display several topics in terms of most probable days given topics,  $\Theta$  (Figure 15) and most probable sequence given topics,  $\Phi$  (Figure 16). We can see that when several hour sequences are discovered, there are often changes in location captured. For example Figure 16 (c), topic 6's most probable sequence is 4411111144444444 with a much higher probability of the last labels (stay region 4's) occurring in the sequence. Note for all users, activity sequences were discovered by the DNTM and we visualize a small set of the most probable topics.



**Fig. 13** The five most probable days given topics for various users. The corresponding sequences learned by the model are in Figure 14. Note that even though we remove the time information from the input sequences, the sequences discovered mostly occur at particular intervals of the day.



**Fig. 14** DNTM results for  $N = 6$ . The most probable sequence for the given user and topic is visualized. The model outputs a distribution of labels given each position in the sequence, which is visualized along the x-axis where the colorbar shows the probability. Note in (b) the sequence is not visible due to a large range between sequence labels (and low probability of the sequence occurring over the entire dataset). However, the most probable days given topics shows the location routine learned.

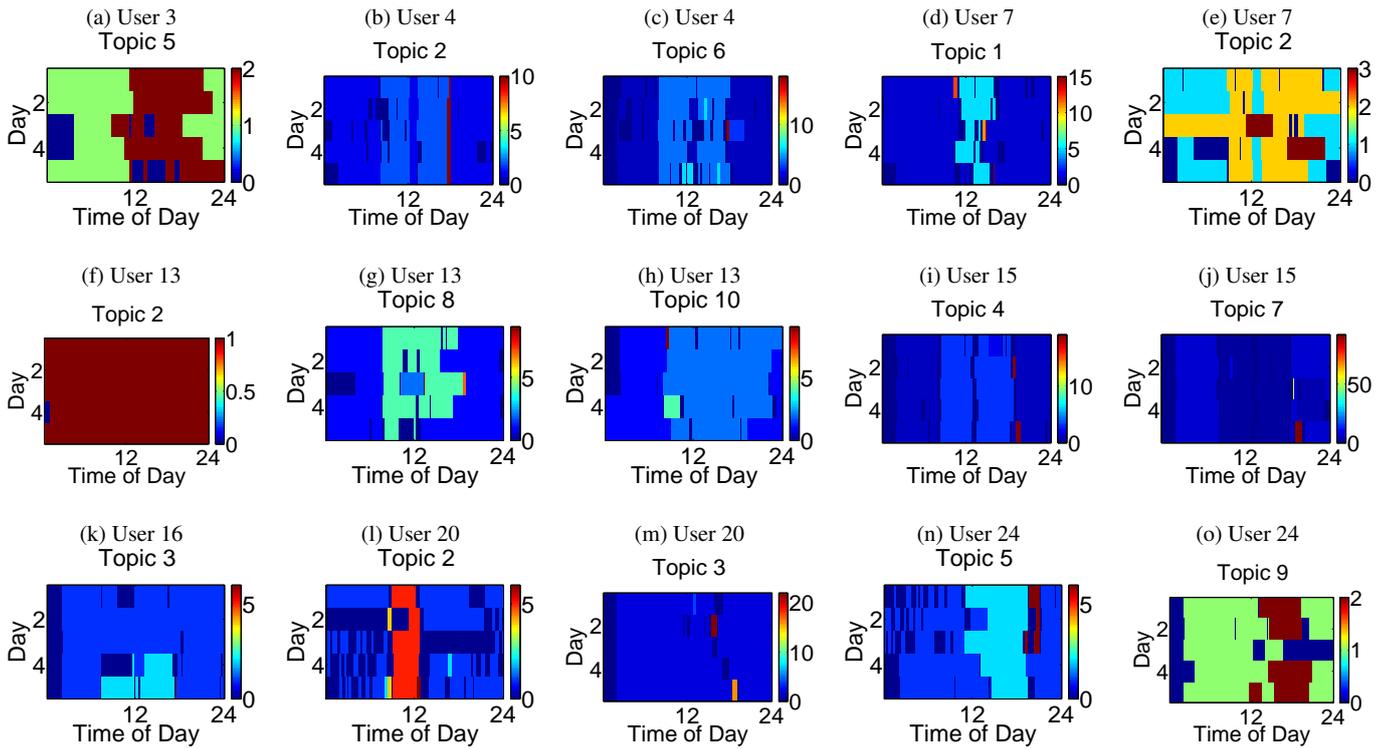
### 6.3 MIT RM Data

For experiments with the MIT dataset, we remove days which contain entirely no reception (N) labels. We experimented with many values of  $T$  and plot selected results for  $T = 20$ . We plot results for the same values of  $\alpha$  and  $\beta$  as in Section 6.1. We consider up to  $N = 14$  corresponding to seven-hour sequences.

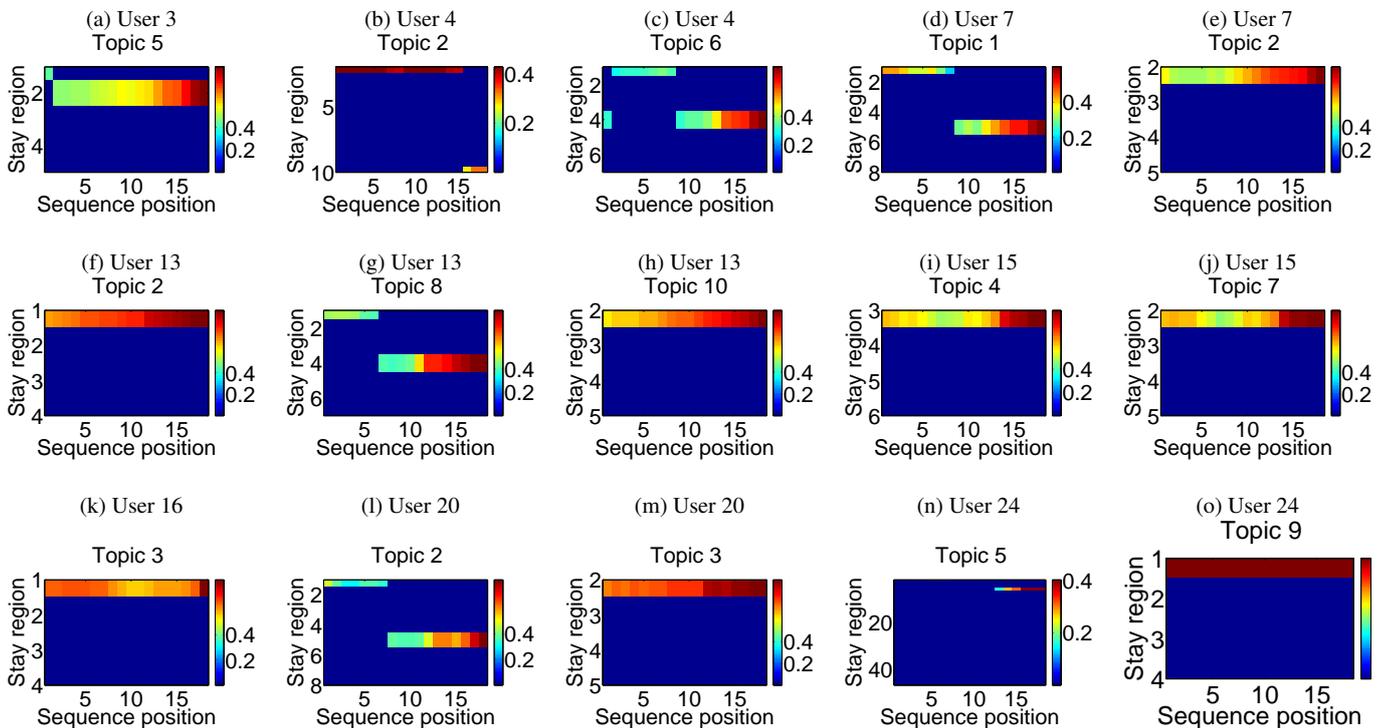
We first visualize a set of 6 topics corresponding to activity sequences for various  $N$ . Note the colorbar indicates the locations. Figure 17 corresponds to dominant sequences discovered for  $N = 3$  (Figure 17 (a)-(c)), and  $N = 13$  (Figure 17 (d)-(f)). We plot the results in terms of the 20 most probable days given topics,  $\theta_m^k$ . The x-axis of the figures corresponds to the time of the day, the y-axis are days, and the legend of the colours are shown to the right of the plots.

In general, we can see emerging location patterns discovered for specific subsets of days in the corpus. For example, in Figure 17 (a) there is 'N' (no reception) in the morning. In (b) there is 'W' (work) after roughly 10 am, with 'O' (out) several hours later, followed by 'W' again. These results resemble the type of results that standard LDA would extract, however, we are able to obtain precise sequence information in our output and "push" the model to output sequences by searching for results at distance  $d$  from the first label in the sequence. As  $N$  increases, we generally discover longer duration location patterns, which are defined in the output parameters of the DNTM model as shown in Tables 2 and 3. Note these tables show the sequences that defined the topics displayed in Figure 17.

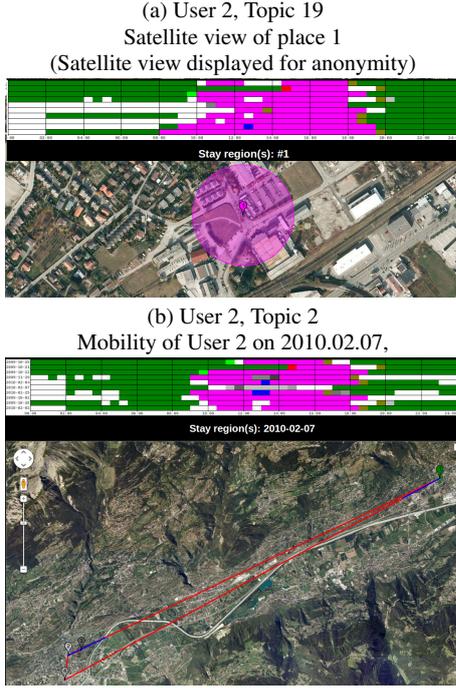
In Table 2, we display the DNTM results in terms of the most probable sequence components given topics. The table



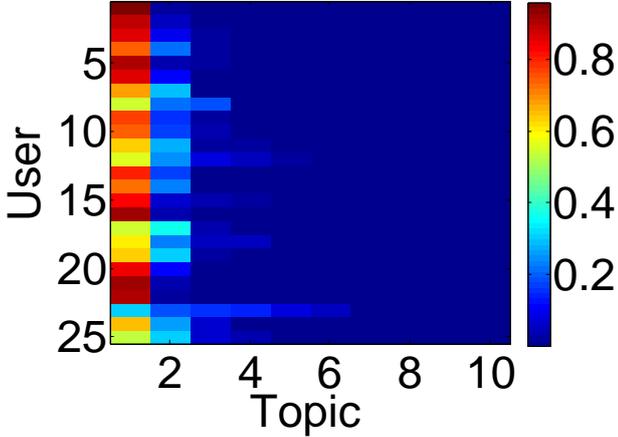
**Fig. 15** The five most probable days given topics for  $N = 18$ . The corresponding sequences learned by the model are in Figure 16. Note that even though we are considering a very long sequence length, the model successfully discovers location behavior patterns.



**Fig. 16** The most probable sequence discovered by the DNTM for  $N = 18$ . Note often the first few most probable sequences discovered by the topic are of interest, but we plot the single most probable sequence for visual clarity.

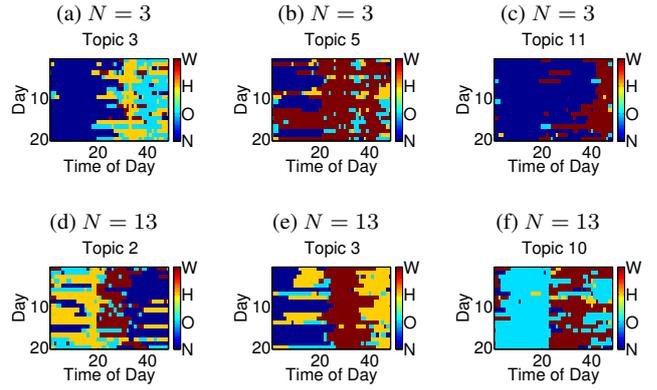


**Fig. 11** Topics and location details for user 2. (a) The satellite view of place 1 is displayed, which corresponds to work for user 2. (b) The mobility for day 2010.02.07 is displayed. The colours of the places displayed on the map correspond to those displayed in the topic. Note that 2010.02.07 is one of the 10 most probable days for user 2 discovered in topic 2 and involved transitions between 3 places.



**Fig. 12** DNTM results over 25 users (y-axis) for 10 topics (x-axis) for  $N = 6$ . The user specific topics are reordered according to their most to least probable topics. For most users a few topics formulate the probability mass.

shows the model output for  $N = 3$ , where the sequence is as follows  $q = (w_1, w_2, w_3)$ . The top ranked sequence components given topics  $k$  are displayed:  $w_2|w_1$  obtained by  $\phi_{2,k}^{(w_1, w_2)_2}$  and  $w_3|w_1$  obtained by  $\phi_{3,k}^{(w_1, w_2)_3}$  along with their probabilities. We do not display  $w_1$  obtained by  $\phi_{1,k}^{w_1}$  since it is inherent in the previous two parameters. We can see the sequence O-O-O starting at 8 pm is discovered in (a)



**Fig. 17** Topics discovered using our model with  $N=3, N=13$ . We plot the results in terms of the 20 most probable days given topics. In general, we can see emerging location patterns discovered within subsets of days in the corpus.

for topic 3 ( $N = 3$ ). The notation '\*' represents any possible location, i.e. O-\*-H indicates that  $w_1 = O$ ,  $w_3 = H$ , with any possible location label for  $w_2$ .

In Table 3 we show the two most probable sequences for the topics displayed in Figure 17(d)-(f). Here, due to the larger value of  $N = 13$ , the actual sequences  $\mathbf{q}$  are displayed. For large  $N$ , we can observe that some of the sequences output are separated in time, for example sequence 2 in (a)  $N = 13$  topic 2. Since we do not force the output to always be a sequence of length  $N$ , there may be more than one sequence of duration less than  $N$  output by the model where the sum of the durations of the sequences output results in  $N$ . Constraints could be imposed to always force length  $N$  sequence as output, though the relaxation of this dependency in our model can be viewed as an advantage. We may in fact be discovering the durations of the dominantly co-occurring sequences. This characteristic is further discussed in the limitations section of the paper. We can see the output obtained by our model contains sequence information, since we obtain probabilities for the labels  $j$  up to distance  $N$  whereas LDA would simply output a probability for each individual label, without any sequence information.

In Figure 18, we plot the perplexity of the DNTM over varying number of topics computed on 20% unseen test data. Note that perplexity is a measure in text modeling of the ability of a model to generalize to unseen data; it is defined as the reciprocal geometric mean of the likelihood of a test corpus given a model. The experiments are conducted for a sequence length of  $N = 8$ . We can see that the perplexity drops to a minimum at around  $T = 50$  topics. We therefore use  $T = 50$  topics in order to compare the performance of our model to LDA. The perplexity results illustrate that for a large number of topics,  $T$ , the model does not seem to overfit the data, since the perplexity does not increase.

**Table 2** Topics discovered using the DNTM corresponding to those displayed in Figure 17, expressed in terms of the most probable sequence components for topics. We show the top ranked sequence components given topics with the probabilities.

(a)  $N = 3$ , Topic 3

$w_2 w_1$	$p(w_2 w_1)$	$w_3 w_1$	$p(w_3 w_1)$
8 pm O-O	0.25	8 pm O-*O	0.23
5 am N-N	0.21	5 am N-*N	0.21

(b)  $N = 3$ , Topic 5

$w_2 w_1$	$p(w_2 w_1)$	$w_3 w_1$	$p(w_3 w_1)$
3:30 pm W-W	0.15	3:30 pm W-*W	0.14
1:30 pm W-W	0.13	1:30 pm W-*W	0.12

(c)  $N = 3$ , Topic 11

$w_2 w_1$	$p(w_2 w_1)$	$w_3 w_1$	$p(w_3 w_1)$
12:30 pm W-W	0.16	12:30 pm W-*W	0.15
5:30 am N-N	0.14	5:30 am N-*N	0.14

**Table 3** Continuation of Table 2. The results in this table are for  $N = 13$  displayed as the sequence  $q$ .

(a)  $N = 13$ , Topic 2

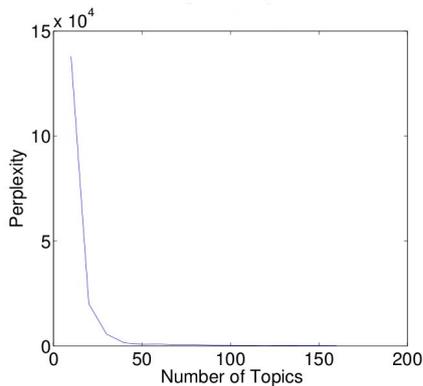
Sequence 1	9 am	H-H-H-H-H-H-W-W-W
Sequence 2	5 pm	N-N-N-N-N
Sequence 2	9 am	H-*-*-*-*W-W-W-W-W

(b)  $N = 13$ , Topic 3

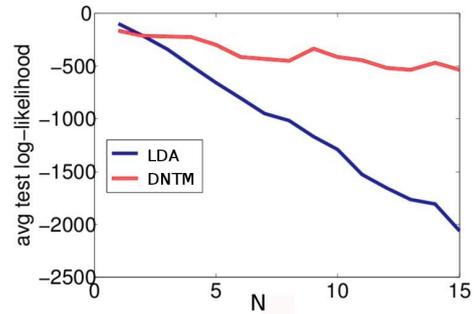
Sequence 1	3 pm	W-W-W-W-W-W-W
Sequence 1	1:30 pm	W-*-*-*-*W-W-W-W-W
Sequence 1	4:30 am	O-*-*-*-*O
Sequence 2	1:30 pm	W-W-W-W-W-W-*-*-*W
Sequence 2	3 pm	W-*-*-*-*W-W-W-W
Sequence 2	4:30 am	O-*-*-*-*O

(c)  $N = 13$ , Topic 10

Sequence 1	4 pm	W-W-W-W-W-W
Sequence 1	4 am	O-*-*-*-*O-O-O-O-O
Sequence 2	4 am	O-O-O-O-O
Sequence 2	4 pm	W-*-*-*-*W
Sequence 2	5 am	O-*-*-*-*O-O-O-O-O



**Fig. 18** Perplexity of the DNTM over the number of topics on 20% unseen days (documents).



**Fig. 19** Average loglikelihood of the DNTM versus LDA on 20% unseen days (documents).

In order to compare our DNTM to LDA, we adapt the vocabulary used for LDA to have a comparable format to that used in the DNTM. The vocabulary we use for LDA consists of a pair of locations, a timeslot, as well as the distance between the locations. This results in a competitive comparison since the key attributes of the DNTM are taken into the vocabulary for LDA. The log-likelihood results on 20% unseen test data, are plotted in Figure 19. We plot the log-likelihood averaged over all the test documents. The log-likelihood results reveal that for small  $N$ , LDA performs slightly better. However, as  $N$  increases, the DNTM consistently has better generalization performance.

## 6.4 Discussion

Though only selected results are presented for the discussion here, many extracted topics correspond to human routines. There are topics corresponding to noise, though they do not dominate the extracted routines.

One evaluation criteria in determining the quality of a model is its predictive power. In Section 6.3 we considered the average loglikelihood of the model on previously unseen data. This is a very general measure giving insight into the predictive capabilities of the model for data that was not previously seen by the model, and the results from Figure 19 are promising for the DNTM.

There are two main limitations of our model. The first one is that there is no constraint forcing the output components to be in sequence. More specifically, a valid output could be  $w_2|w_1, z$  and  $w_3|w_1, z$  where  $w_1 \neq w_1'$ . In our experiments, we found that this effect did not occur often in the output. This can also be an advantage in that the output generates varying length sequences and determines the actual sequence lengths of the activities since they may not necessarily be  $N$ . We would have to add some constraints to the model in order to always force the output to be sequences of length  $N$ . Another limitation is that the output can contain overlapping components. For example, using the data from Section 6.3, a valid sequence output for a topic may be 3:30 pm H-H and 3 pm H-\*H. Here, the sequence output is

not of length 3. To address this problem, again, some constraints should be imposed regarding the time component in the feature construction.

## 7 Conclusions

In this paper we proposed the distant n-gram topic model as an alternative to model long sequences for activity modeling, and apply it in the context of human location sequences. Considering two real life human datasets collected via mobile phone location logs, we tested our model firstly on locations obtained by smartphones based on GPS and wifi, and secondly by cell tower location features. The patterns extracted by our model are meaningful and are further validated by considering a synthetic dataset. We evaluated our model against LDA considering log-likelihood performance on unseen data and found that the DNTM outperforms LDA for most of the studied cases.

There are several future directions for this work. The first direction is to explore extensions of the proposed model. One could extend the DNTM by taking into account the limitations mentioned and imposing application-specific constraints. One can also further investigate the dependence problem and consider methods to model dependence among labels as opposed to always having the label dependent on the first element, though this could quickly lead to parameter size explosion. For example, there may be effective hierarchical methods for determining the number of previous labels that a given label in a sequence should depend on. The second direction of extensions would be to consider other types of data, for example in the context of other wearable data and activities. Finally, one other relevant line of work future work is a comparison of our method with Hidden Markov Models learned in an unsupervised setting, imposing structure to learn long-term sequential patterns.

## Acknowledgements

This research was funded by the SNSF HAI project and the LS-CONTEXT project funded by Nokia Research. K. Farrahi also acknowledges the Socionical project and the Pervasive Computing Group at JKU, Linz. We thank Olivier Bornet (Idiap) for help with location data processing and visualization, Gian Paolo Perrucci (Nokia Research Lausanne) for insights on routine visualization, and Trinh-Minh-Tri Do (Idiap) for discussions on sequence modeling methods.

## References

1. Ashbrook, D., Starner, T.: Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing* **7**(5), 275–286 (2003)
2. Bao, T., Cao, H., Chen, E., Tian, J., Xiong, H.: An unsupervised approach to modeling personalized contexts of mobile users. In: IEEE International Conference on Data Mining (ICDM), pp. 38–47 (2010)
3. Becker, R.A., Cáceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C.: Route classification using cellular handoff patterns. In: *UbiComp*, pp. 123–132 (2011)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. *Journal of Machine Learning Research* **3**, 993–1022 (2003)
5. Candia, J., Gonzalez, M.C., Wang, P., Schoenharl, T., Madey, G., Barabasi, A.L.: Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical* **41**(22), 224,015+ (2008)
6. Do, T., Gatica-Perez, D.: Groupus: Smartphone proximity data and human interaction type mining. In: *Proc. IEEE Int.Symp. on Wearable Computers (ISWC)*. San Francisco, USA (2011)
7. Eagle, N., Pentland, A.: Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology* **63**(7), 1057–1066 (2009)
8. Farrahi, K.: A probabilistic approach to socio-geographic reality mining. Ph.D. thesis, Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland (2011). DOI 10.5075/epfl-thesis-5018. URL <http://library.epfl.ch/theses/?nr=5018>
9. Farrahi, K., Gatica-Perez, D.: Mining human location-routines using a multi-level topic model. In: *Socialcom Symposium on Social Intelligence and Networking (Socialcom SIN)*. Minneapolis, USA (2010)
10. Farrahi, K., Gatica-Perez, D.: Probabilistic mining of socio-geographic routines from mobile phone data. *IEEE Journal of Selected Topics in Signal Processing (J-STSP)* **4**(4), 746–755 (2010)
11. Farrahi, K., Gatica-Perez, D.: Extracting mobile behavioral patterns with the distant n-gram topic model. In: *International Symposium on Wearable Computers (ISWC)*, pp. 1–8 (2012)
12. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. *Nature* **453**(7196), 779–782 (2008)
13. Görnerup, O.: Scalable mining of common routes in mobile communication network traffic data. In: *Pervasive*, pp. 99–106. Newcastle upon Tyne, UK (2012)
14. Griffiths, T.L., Steyvers, M.: Finding scientific topics. *Proceedings of the National Academy of Sciences* **101**(Suppl. 1), 5228–5235 (2004)
15. Hightower, J., Consolvo, S., Lamarca, A., Smith, I., Hughes, J.: Learning and recognizing the places we go. In: *Ubiquitous Computing (UbiComp)*, pp. 159–176. Tokyo, Japan (2005)
16. Hofmann, T.: Probabilistic latent semantic analysis. In: *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 289–296. Stockholm, Sweden (1999)
17. Huynh, T., Fritz, M., Schiele, B.: Discovery of activity patterns using topic models. In: *Ubiquitous computing (UbiComp)*, pp. 10–19. Seoul, Korea (2008)
18. Kang, J.H., Welbourne, W., Stewart, B., Borriello, G.: Extracting places from traces of locations. *ACM SIGMOBILE Mobile Computing and Communications Review* **9**(3), 58–68 (2005)
19. Kiukkonen, N., Blom, J., Dousse, O., Gatica-Perez, D., Laurila, J.: Towards rich mobile phone datasets: Lausanne data collection campaign. In: *Proc. ACM Int. Conf. on Pervasive Services (ICPS)*. Berlin, Germany (2010)
20. Liao, L., Fox, D., Kautz, H.: Location-based activity recognition. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 787–794. Vancouver, Canada (2006)
21. Mackay, D.J.C.: *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press (2003)
22. Marmasse, N., Schmandt, C.: Location-aware information delivery with commotion. pp. 157–171. Springer (2000)
23. Montoliu, R., Gatica-Perez, D.: Discovering human places of interest from multimodal mobile phone data. In: *Proc. ACM Int. Conf. on Mobile and Ubiquitous Multimedia (MUM)*. Limassol, Cypress (2010)

24. Patterson, D., Liao, L., Fox, D., Kautz, H.: Inferring high-level behavior from low-level sensors. In: Ubiquitous Computing (UbiComp), pp. 73–89. Seattle, USA (2003)
25. Petterson, J., Smola, A.J., Caetano, T.S., Buntine, W.L., Narayana-murthy, S.: Word features for latent dirichlet allocation. In: Advances in Neural Information Processing Systems (NIPS), pp. 1921–1929 (2010)
26. Phithakkitnukoon, S., Horanont, T., Lorenzo, G.D., Shibasaki, R., Ratti, C.: Activity-aware map: Identifying human daily activity pattern using mobile phone data. In: Proceedings of the First international conference on Human behavior understanding, pp. 14–25. Springer-Verlag, Berlin, Heidelberg (2010)
27. Varadarajan, J., Emonet, R., Odobez, J.M.: Sparsity in topic models. In: Practical Applications of Sparse Modeling: Biology, Signal Processing and Beyond. MIT Press (2012)
28. Wallach, H.: Topic modeling: beyond bag-of-words. In: Proc. of the International Conference on Machine Learning (ICML). Pittsburgh, USA (2006)
29. Wang, X., McCallum, A., Wei, X.: Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: IEEE International Conference on Data Mining (ICDM), pp. 697–702. Washington, USA (2007)
30. Yavas, G., Katsaros, D., Ulusoy, O., Manolopoulos, Y.: A data mining approach for location prediction in mobile environments. Data Knowl. Eng. **54**(2), 121–146 (2005)
31. Zheng, J., Ni, L.M.: An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data. In: Ubiquitous Computing (UbiComp). Pittsburgh, USA (2012)
32. Zheng, Y., Zhang, L., Xie, X., Ma, W.Y.: Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th international conference on World wide web, pp. 791–800. ACM, New York, NY, USA (2009)

## Appendix: derivation of the distant n-gram topic model

From the graphical model in Figure 1, we can determine the following relationship:

$$\begin{aligned}
p(\mathbf{z}, \mathbf{q} | \alpha, \beta) &= p(\mathbf{z}, \mathbf{w}_1, \dots, \mathbf{w}_N | \alpha, \beta) \\
&= p(\mathbf{w}_1, \dots, \mathbf{w}_n | \mathbf{z}, \alpha, \beta) \cdot p(\mathbf{z} | \alpha, \beta) \\
&= p(\mathbf{w}_2, \dots, \mathbf{w}_n | \mathbf{w}_1, \mathbf{z}, \alpha, \beta) \cdot p(\mathbf{w}_1 | \mathbf{z}, \alpha, \beta) \cdot p(\mathbf{z} | \alpha, \beta) \\
&= p(\mathbf{z} | \alpha) p(\mathbf{w}_1 | \mathbf{z}, \beta_1) \prod_{j=2}^N p(\mathbf{w}_j | \mathbf{z}, \mathbf{w}_1, \beta_j) \\
&= \int_{\Theta} p(\mathbf{z} | \Theta) p(\Theta | \alpha) d\Theta \cdot \int_{\Phi_1} p(\mathbf{w}_1 | \mathbf{z}, \Phi_1) p(\Phi_1 | \beta_1) d\Phi_1 \cdot \\
&\quad \prod_{j=2}^n \int_{\Phi_j} p(\mathbf{w}_j | \mathbf{w}_1, \mathbf{z}, \Phi_j) p(\Phi_j | \beta_j) d\Phi_j \\
&= \prod_{m=1}^M \left( \frac{1}{B(\alpha)} \int \prod_{k=1}^T \theta_{m,k}^{n_m^k + \alpha - 1} d\theta \right) \cdot \prod_{k=1}^T \left( \frac{1}{B(\beta_1)} \int \prod_{t=1}^V \phi_{1,k,t}^{n_k^t + \beta_1 - 1} d\phi_1 \right) \cdot \\
&\quad \prod_{j=2}^n \prod_{k=1}^T \frac{1}{B(\beta_j)} \left( \int \prod_{t_1=1}^V \prod_{t_2=1}^V \phi_{j,k,t_1,t_2}^{n_{k_i}^{(t_1,t_2)j} + \beta_j - 1} d\phi_j \right) \\
&= \prod_{m=1}^M \frac{B(n_m + \alpha)}{B(\alpha)} \cdot \prod_{k=1}^T \left( \frac{B(n_k + \beta_1)}{B(\beta_1)} \right) \cdot \prod_{j=2}^n \frac{B(n_{k_j} + \beta_j)}{B(\beta_j)}
\end{aligned}$$

The joint probability of observations and latent topics can be obtained by marginalizing over the hidden parameters  $\Theta$  and  $\Phi$ . These relations are then used for inference and parameter estimation where  $p(\mathbf{z} | \alpha)$ ,  $p(\mathbf{w}_1 | \mathbf{z}, \beta_1)$ , and  $p(\mathbf{w}_j | \mathbf{w}_1, \mathbf{z}, \beta_j)$  are derived in [8] resulting in the following.

$$p(\mathbf{z} | \alpha) = \prod_{m=1}^M \frac{B(n_m + \alpha)}{B(\alpha)} \quad (11)$$

$$p(\mathbf{w}_1 | \mathbf{z}, \beta_1) = \prod_{k=1}^T \frac{B(n_k + \beta_1)}{B(\beta_1)} \quad (12)$$

$$p(\mathbf{w}_j | \mathbf{w}_1, \mathbf{z}, \beta_j) = \prod_{k=1}^T \frac{B(n_{k_j} + \beta_j)}{B(\beta_j)}, 1 < j \leq n \quad (13)$$

We then derive the model parameters based on the MCMC approach of Gibbs sampling [14].

$$p(z_i = k | \mathbf{z}_{-i}, \mathbf{q}, \alpha, \beta) = \frac{p(\mathbf{z}, \mathbf{q} | \alpha, \beta)}{p(\mathbf{z}_{-i}, \mathbf{q} | \alpha, \beta)} \quad (14)$$

using the knowledge  $\mathbf{z}_{-i}$ , or  $\mathbf{w}_{x_{-i}}$  indicate that token  $i$  is excluded from the topic or label  $\mathbf{w}_x$

$$\propto \frac{B(n_m + \alpha)}{B(n_{m-i} + \alpha)} \cdot \frac{B(n_k + \beta_1)}{B(n_{k-i} + \beta_1)} \cdot \prod_{j=2}^n \frac{B(n_{k'_j} + \beta_j)}{B(n_{k'_{j,-i}} + \beta_j)} \quad (15)$$

Note the proportionality stems from the terms  $\mathbf{w}_{1_i}$  and  $\mathbf{w}_{j_i}$

$$\propto (n_{m,-i}^k + \alpha) \cdot \frac{n_{k,-i}^{w_1} + \beta_1}{\sum_{w_1=1}^V n_{k,-i}^t + \beta_1} \quad (16)$$

$$\prod_{j=2}^n \frac{n_{k,-i}^{(w_1, w_2)_j} + \beta_j}{\sum_{w_1=1}^V \sum_{w_2=1}^V n_{k,-i}^{(w_1, w_2)_j} + \beta_j}$$

where  $n_x^{(y)} = n_{x,-i}^{(y)} + 1$  if  $x = x_i$  and  $y = y_i$

and  $n_x^{(y)} = n_{x,-i}^{(y)}$  in other cases.

where  $n_k = \{n_k^{w_1}\}_{w_1=1}^V$  and  $n_{k'_j} = \{n_{k'_j}^{(w_1, w_2)_j}\}_{w_1=1, w_2=1}^{w_1=V, w_2=V}$ .

We use the properties  $B(x) = \frac{\prod_{k=1}^{dim x} \Gamma(x_k)}{\Gamma(\sum_{k=1}^{dim x} x_k)}$ , and  $\Gamma(y) = (y-1)!$ .

The model parameters can then be estimated as follows:

$$\theta_m^k = \frac{n_m^k + \alpha}{\sum_{k=1}^T (n_m^k + \alpha)} \quad (17)$$

$$\phi_{1,k}^t = \frac{n_k^t + \beta_1}{\sum_{w_1=1}^V (n_k^{w_1} + \beta_1)} \quad (18)$$

$$\phi_{j,k}^{(w_1, w_2)_j} = \frac{n_k^{(w_1, w_2)_j} + \beta_j}{\sum_{w_1=1}^V \sum_{w_2=1}^V (n_k^{(w_1, w_2)_j} + \beta_j)} \quad (19)$$