TOPTRAC: Topical Trajectory Pattern Mining

Younghoon Kim*
Hanyang Univ.
Ansan, Korea
younghoonk79@gmail.com

Jiawei Han
UIUC
Illinois, US
hanj@illinois.edu

Cangzhou Yuan*
Beihang Univ.
Beijing, China
yuancz@buaa.edu.cn

ABSTRACT

With the increasing use of GPS-enabled mobile phones, geo-tagging, which refers to adding GPS information to media such as micro-blogging messages or photos, has seen a surge in popularity recently. This enables us to not only browse information based on locations, but also discover patterns in the location-based behaviors of users. Many techniques have been developed to find the patterns of people's movements using GPS data, but latent topics in text messages posted with local contexts have not been utilized effectively.

In this paper, we present a latent topic-based clustering algorithm to discover patterns in the trajectories of geo-tagged text messages. We propose a novel probabilistic model to capture the semantic regions where people post messages with a coherent topic as well as the patterns of movement between the semantic regions. Based on the model, we develop an efficient inference algorithm to calculate model parameters. By exploiting the estimated model, we next devise a clustering algorithm to find the significant movement patterns that appear frequently in data. Our experiments on real-life data sets show that the proposed algorithm finds diverse and interesting trajectory patterns and identifies the semantic regions in a finer granularity than the traditional geographical clustering methods.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Spatial databases and GIS

General Terms

Algorithms

Keywords

Topical trajectory pattern; modeling geo-tagged messages

*This work was done when Y. Kim and C. Yuan were visiting University of Illinois at Urbana-Champaign.

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2783342

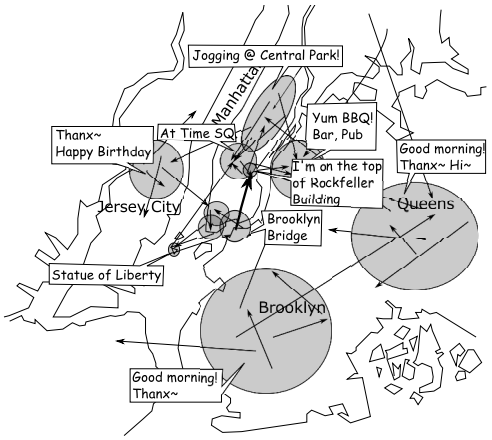


Figure 1: Trajectories in New York City

1. INTRODUCTION

Micro-blogging services such as Twitter and Tumblr have emerged recently as a medium in spotlight for communication, as we can update our status anywhere and anytime by using a mobile device. With the increasing use of GPS-enabled mobile phones, user-generated media with a geo-tag such as text messages and photos have been produced with an accelerating trend in the micro-blogging services. This enables us to not only retrieve information with locations, but also utilize the patterns of location-based user behaviors.

Studies on the trajectory data collected using GPS sensors have always been important topics for many location-based applications such as mobile navigation system[5], itinerary suggestion[19], urban traffic analysis[4], and tracking of hurricanes or animals[11]. However, these traditional trajectory mining techniques assume that the GPS sensors record the location of moving objects frequently. They also do not consider the semantic meanings of the locations which are available in the associated tags and short text messages on social media services.

Recently, trajectory mining techniques to use *low-sampling semantic* trajectory data collected from social media services have been developed to recommend travel routes[10] or find interesting trajectory patterns[16, 18]. However, since they assume that well-refined 'regions of interest' and 'semantic labels' such as user-generated categories are available, it is still hard to utilize those techniques to discover semantic trajectory patterns with respect to 'latent topics' of geo-tagged messages posted in the local context.

Trajectory mining with geo-tagged messages posted in social media has to deal with three difficulties: (1) how to find topically coherent regions, (2) how to handle noisy messages such as everyday conversations and (3) how to overcome the sparsity of trajectory patterns. Since geo-tag is a pair of real numbers which denote latitude and longitude, we first need to cluster geo-tagged messages posted at the close site with a similar topic. Let us call such regions with a coherent topic preference *semantic regions*. Because most of the messages posted in micro-blogging services are personal and ordinary talks, it is very difficult to cluster them by the coherence of topics. Consequently, it becomes hard to identify the trajectory patterns with useful information out of noisy trajectories.

Consider the geo-tagged messages posted at New York city as shown in Figure 1. We can expect that there are popular places where messages should be posted with a coherent topic such as Time Square, Central Park and Statue of Liberty. However, even in such regions, many messages are personal without any local context such as 'Good morning' or 'Happy birthday'. Such noisy messages make it hard to identify either topical semantic regions or trajectory patterns. Furthermore, suppose many travelers move from the popular attractions such as Brooklyn Bridge to Rockfeller Building as posting geo-tagged messages. However, since more people visit the neighborhood, such as Time Square, and post messages with various reasons, the topic of Rockfeller Building becomes hard to find and the pattern from Brooklyn Bridge to Rockfeller Building may also be missed.

To deal with these problems, we develop a trajectory clustering algorithm, called *TOPTRAC*, which is **TOP**ical **TRA** je**C**tory pattern mining in this paper. Our algorithm is designed to discover topically coherent regions with a finer granularity by weighting the messages whose previous one also was posted in a semantic region. For example, even though the messages on Rockfeller Building are relatively sparser than those on Time Square, if there is a transition pattern of people who have arrived from another semantic region such as Brooklyn Bridge, we can identify the semantic region of Rockfeller Building. Furthermore, this also helps us discover various trajectory patterns by clustering semantic regions with a finer granularity.

To discover semantic regions with a coherent topic, Yin et al. have suggested a clustering algorithm, called LGTA, in [17]. However, since it focused on finding topically coherent locations only, it cannot cope with the other two difficulties.

To the best of our knowledge, our algorithms presented here are the first work for the problem of topical trajectory pattern mining. The contributions of this paper are summarized as follows:

- We introduce a new topical clustering problem in the trajectories of geo-tagged text messages.
- We propose a probabilistic model to discover not only the latent semantic regions where people post messages with a coherent topic but also the diverse patterns of movements between the semantic regions based on the observation that people may visit a place with various purposes.
- We develop an efficient variational EM algorithm for posterior inference and parameter estimation.
- To find significant patterns using the estimated model, we devise a dynamic programming algorithm to compute the most likely sequence of latent semantic regions for a given sequence, which works similar to the Viterbi algorithm of the hidden Markov model.
- In experiments, we show that our method finds useful semantic regions in a fine granularity as well as diverse movement patterns.

The rest of paper is structured as follows: Section 2 presents the related works of trajectory pattern mining. In Section 4.1, we define the notations to use in this paper and introduce our generative model. We next develop a dynamic programming algorithm to identify trajectory patterns in Section 4.4 and present experimental results in Section 5. Finally, we conclude our study in Section 6.

2. RELATED WORK

Trajectory mining with GPS data has been extensively studied to find patterns in the movements of objects. We first discuss about the traditional trajectory mining techniques[3, 11, 19] which utilize GPS locations only and next study the recent algorithms[7, 10, 16, 17, 18] that find trajectory patterns using the semantics of location.

Traditional trajectory mining algorithms usually cluster the similar transactions to find common and popular trajectories. For example, Lee et al. proposed a clustering algorithm in [11] which discovers not only the groups of similar trajectories but also the clusters with similar sub-sequences. Zheng et al. introduced a ranking method in [19] based on a HIT algorithm to find popular trajectories. Furthermore, an algorithm finding the most popular path between given two locations was presented in [3]. However, these traditional trajectory mining typically have focused on the transaction data obtained from the GPS sensors which records locations very frequently, while geo-tagged messages in micro-blogging services are usually posted sparsely.

To discover the semantic patterns from the GPS locations tagged in pictures, the algorithm proposed in [10] finds sequential patterns of user-generated tags. Since long trajectory patterns rarely exist in the trajectory of such a social media, Yin et al.[16] investigated the problem of finding diverse and interesting trajectory patterns. Recently, Zhang et al.[18] presented a clustering algorithm to find trajectory patterns that are divided delicately by utilizing the user-generated categories of locations. However, since these techniques assume that refined tags which represent the semantic of point of interests are provided, it is hard for us to simply modify for finding topical transition patterns from the geo-tagged messages.

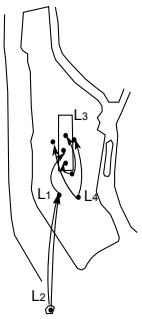
To find topics associated with GPS locations from the geo-tagged messages, Yin et al.[17] suggested a probabilistic model-based clustering algorithm, called LGTA, that groups the geo-tagged messages which have not only similar locations but also coherent topics. In [7] and [8], extended probabilistic models of LGTA was developed to capture the patterns of individual users. Since LGTA in [17] can be simply extended for our problem of topical transition pattern mining, we compare our proposed algorithm to the extended LGTA which is presented in Appendix A.

3. PRELIMINARIES

In this section, we first provide the definitions to be used in the rest of the paper and then formulate our problem of topical trajectory pattern mining.

	(message, location)
s_1	$m_{1,1}$=(Start today's tour, L_2) $m_{1,2}$=(Time square, L_1)
s_2	$m_{2,1}$=(Apple store, L_5) $m_{2,2}$=(Running in Central park, L_3)
s_3	$m_{3,1}$=(Many people @ Time SQ, L_1) $m_{3,2}$=(Central park zoo, L_3)
s_4	$m_{4,1}$=(Liberty Statue tour, L_2) $m_{4,2}$=(I'm at Time square, L_1)
s_5	$m_{5,1}$=(Now, MoMA, L_4) $m_{5,2}$=(Metro. museum of art, L_3)
s_6	$m_{6,1}$=(At MoMA, L_4) $m_{6,2}$=(Museum of natural history, L_3)
	...

(a) Trajectories



(b) A map

Figure 2: An example of geo-tagged messages

3.1 Notations for Observed Data

Here we define the notations to describe the collection of geo-tagged messages posted in micro-blogging services such as Twitter.

DEFINITION 3.1.: *A* trajectory *is the sequence of geo-tagged messages posted by a user in a given time interval such as a day or a week, where the messages are listed in the order of the posting time.*

Let $\mathbb{C}=\{s_1, \ldots, s_N\}$ be a collection of N trajectories where each trajectory is denoted by s_t with t=1,...,N. We also let $\mathcal{V}=\{1, \ldots, V\}$ be a set of word IDs in which each word appears at least once in the collection. A trajectory $s_t=\langle m_{t,1:N_t}\rangle$ consists of N_t geo-tagged messages where the i-th message in s_t is represented by $m_{t,i}$. We use $m_{1:N}$ to denote a sequence of variables m_i (i=1, ..., N).

Each geo-tagged message $m_{t,i}$ consists of a geo-tag $\vec{G}_{t,i}$ and a bag-of-words $\mathbf{w}_{t,i}$ such that

- $\vec{G}_{t,i}$ is a 2-dimensional vector $(G_{t,i,x}, G_{t,i,y})$ representing the latitude and longitude respectively,
- $\mathbf{w}_{t,i}$ is a bag-of-words which contains $N_{t,i}$ words $\{w_{t,i,1}, \ldots, w_{t,i,N_{t,i}}\}$ where $w_{t,i,j}$ denotes the j-th word in $\mathbf{w}_{t,i}$ and should be one of the word IDs in the vocabulary $\mathcal{V}$ (i.e., $w_{t,i,j} \in \mathcal{V}$).

Furthermore, we use $m_{t,i}.timestamp$ to denote the time when $m_{t,i}$ was posted.

3.2 Problem Formulation

We first define the *latent semantic region* and *topical transition pattern* as follows.

DEFINITION 3.2.: *A* latent semantic region *or* semantic region *is a geographical location where geo-tagged messages are posted with the same topic preference.*

DEFINITION 3.3.: *A* topical transition pattern *or* transition pattern *is a movement from one semantic region to another frequently in the collection $\mathbb{C}$ of trajectories. Furthermore, for each transition pattern, we refer to a pair of actual geo-tagged messages, which illustrate the topics of semantic regions in the transition pattern, as the* transition snippets *of the pattern.*

We focus on the discovery of the trajectory patterns with length 2 only in our paper since longer frequent trajectory patterns rarely exist in low-sampling trajectories as discussed in [16]. We now formulate the problem as follows:

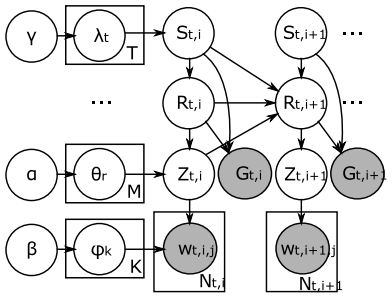


Figure 3: Graphical presentation of our model

DEFINITION 3.4.: **Topical trajectory mining problem***: Given a collection $\mathbb{C}$ of geo-tagged message trajectories,* topical trajectory mining *is to find topical transition patterns and the top-k transition snippets which best represent each transition pattern.*

EXAMPLE 3.5.: *Consider the six sequences of geo-tagged messages in Figure 2. Each sequence contains two messages as shown in Figure 2(a) and the locations where the messages are posted are plotted in Figure 2(b). A look at the sequences s_1 and s_4 suggests that there is a transition pattern from 'Statue of Liberty' to 'Time Square', whose regions are represented as L_2 and L_1 respectively. Furthermore, the pairs of geo-tagged messages $\langle m_{1,1}, m_{1,2}\rangle$ in s_1 and $\langle m_{4,1}, m_{4,2}\rangle$ in s_4 can be regarded as the transition snippets of the pattern.*∎

4. FINDING TOPICAL TRAJECTORY PATTERNS

Our proposed algorithm is organized as follows:

(1) In Section 4.1, we first present our novel probabilistic model to describe the generative process of posting geo-tagged messages in Twitter. Then, we provide an inference algorithm to learn the latent semantic regions and the transition probabilities between regions.

(2) The probabilistic model enables us to identify the latent semantic regions and compute the probabilities with which a given geo-tagged message is posted at those regions. Thus, in Section 4.4, for each geo-tagged message in the collected sequences, we find the most probable semantic region where each message was posted at using the calculated probabilities.

(3) In the following Section 4.5, we present an algorithm to find transition patterns that frequently occur in the collected messages as well as the method to discover the top-k representative geo-tagged messages which best show the frequent transition patterns.

4.1 Generative Model

In our probabilistic model, each sequence of geo-tagged messages are generated independently and identically by a user. However, each geo-tagged message in a sequence is produced depending on *its preceding message* in the sequence.

Generative process for geo-tagged messages: We assume that there are **M** latent semantic regions and **K** hidden topics in the collection $\mathbb{C}$ of geo-tagged messages. The topic distribution for a semantic region r is denoted by $\vec{\theta}_r=\langle\theta_{r,1}, \ldots, \theta_{r,K}\rangle$, where $\theta_{r,k}$ with $1\leq k\leq K$ is the probability that the topic k is chosen to generate a text message

	Description	Approx. vars.
$\vec{\theta}_r$	Topic distribution in region r	$q(\vec{\theta}_r)$=Dir($\vec{a}_r$)
$\vec{\phi}_k$	Word distribution for the k-th topic	$q(\vec{\phi}_k)$=Dir($\vec{b}_k$)
$\vec{\lambda}_t$	Bernoulli distribution for the relationship to the local context in s_t	$q(\vec{\lambda}_t)$=Beta($\vec{c}_t$)
$S_{t,i}$	Random variable representing whether $m_{t,i}$ is in the local context or not	$q(S_{t,i})$ $=\langle\sigma_{t,i,0}, \sigma_{t,i,1}\rangle$
$R_{t,i}$	Random variable for the latent semantic region of $m_{t,i}$	$q(R_{t,i})$ $=\langle\rho_{t,i,1}, .., \rho_{t,i,M}\rangle$
$Z_{t,i}$	Random variable that indicates the latent topic used to generate $m_{t,i}$	$q(Z_{t,i})$ $=\langle\zeta_{t,i,1}, .., \zeta_{t,i,K}\rangle$
$\vec{\delta}$	Probability distribution of selecting a starting latent semantic region	-
$\vec{\delta}_{r,k}$	Transition probability distribution from region r with topic k	-
$\vec{\mu}_r$, Σ_r	Mean and covariance matrix of r	-
$\vec{\alpha}$,$\vec{\beta}$,$\vec{\gamma}$	Hyper-parameters for the Dirichlet priors of $\vec{\theta}_r$, $\vec{\phi}_k$ and $\vec{\lambda}_t$ respectively	-

Figure 4: The variables of our model

in the sematic region r. Similarly to the LDA model[2], we represent the distribution over the vocabulary $\mathcal{V}$ of a topic k as $\vec{\phi}_k$=$\langle\phi_{k,1}, \ldots, \phi_{k,V}\rangle$, where $\phi_{k,t}$ denotes the probability that a word t is selected for the topic k.

The difference between the LDA model and our model is that we select only a topic for each text message and choose every word in the message according to the word distribution of the selected topic repeatedly. This assumption has been widely adopted in many Bayesian models for Twitter such as [7, 8] because the length of a message is usually very short due to the 140 character limitation.

Let $R_{t,i}$ and $Z_{t,i}$ denote the random variables of the latent semantic region and latent topic selected for the message $m_{t,i}$ in s_t respectively. For each sequence $s_t \in \mathbb{C}$, we assume that there is a Bernoulli distribution $\vec{\lambda}_t$ which decides whether each geo-tagged message in s_t is related to the semantic region or not (i.e., the message is posted in a local context or not.)

(1) If a message $m_{t,i}$ does not belong to any latent semantic region (i.e., a location-irrelevant message), the region $R_{t,i}$ of $m_{t,i}$ is selected as any one among the M regions with probability $1/M$ and its geo-tag is also generated with uniform probability f_0. If $m_{t,i}$ is posted regardless of the local context, we denote the event with the random variable $S_{t,i}$=0, otherwise, $S_{t,i}$=1.

(2) If $m_{t,i}$ has a local context (i.e., $S_{t,i}$=1), we select a latent semantic region $R_{t,i}$ depending on whether its preceding message is related to a semantic region or not (i.e., $S_{t,i-1}$=1 or 0), and if $S_{t,i-1}$=1, $R_{t,i}$ is chosen depending on which latent region and topic are involved to generate the previous message $m_{t,i-1}$.

In details, for case (2), if the message $m_{t,i}$ is the first in the sequence (i.e., i=1) or the previous geo-tagged message is not related to a local context (i.e., $S_{t,i-1}$=0), we select a region $R_{t,i}$ following a categorical distribution over M latent semantic regions, denoted by $\vec{\delta}_0$, independently to its previous message. If its preceding message $m_{t,i-1}$ is also generated in a local context (i.e., $S_{t,i-1}$=1), $R_{t,i}$ is chosen following a categorical distribution $\vec{\delta}_{R_{t,i-1},Z_{t,i-1}}$=$\langle\delta_{R_{t,i-1},Z_{t,i-1},1}$, $\ldots, \delta_{R_{t,i-1},Z_{t,i-1},M}\rangle$ where $\delta_{R_{t,i-1},Z_{t,i-1},r}$ denotes the transition probability with which a user submits a message in the region r after posting a message in $R_{t,i-1}$ on the topic $Z_{t,i-1}$.

With a selected region $R_{t,i}$, we choose a topic $Z_{t,i}$ following $\vec{\theta}_{R_{t,i}}$ and select every word in $\mathbf{w}_{t,i}$ following $\vec{\phi}_{Z_{t,i}}$. Furthermore, we select a geo-tag $\vec{G}_{t,i} \in \mathbb{R}^2$ according to a 2-dimensional Gaussian probability function:

$$f_{R_{t,i}}(\vec{G}_{t,i}) = \frac{1}{2\pi\sqrt{|\Sigma_{R_{t,i}}|}} \exp\left(-\frac{1}{2}(\vec{G}_{t,i} - \vec{\mu}_{R_{t,i}})^\top \Sigma^{-1}_{R_{t,i}}(\vec{G}_{t,i} - \vec{\mu}_{R_{t,i}})\right) \quad (1)$$

where $\vec{\mu}_R$ and Σ_R represent the mean and covariance matrix of a latent semantic region R's Gaussian distribution. In Figure 4, we list the notations of posterior distributions and model parameters used in our generative model. Furthermore, we summarize the above generative process in Figure 5 and the graphical representation of our model is presented in Figure 3.

Aims of our model: By introducing the random variable $S_{t,i}$ to determine the relation to a local context and using a uniform distribution to generate geo-tags for the messages with $S_{t,i}$=0, we can identify local-irrelevant messages appearing everywhere. Furthermore, the use of transition probability $\vec{\delta}_{r,k}$ enables us to weight a semantic region which has a high conditional probability that people move into the region after visiting another semantic region.

For each region r=1, ..., M:
- Select a categorical distribution: $\vec{\theta}_r \sim$ Dir($\vec{\alpha}$)

For each topic k=1, ..., K:
- Select a categorical distribution: $\vec{\phi}_k \sim$ Dir($\vec{\beta}$)

For each sequence s_t=$\langle m_{t,1}, \ldots, m_{t,N_t}\rangle \in \mathbb{C}$:
- Select a Bernoulli distribution: $\lambda_t \sim$ Beta($\vec{\gamma}$)
- For each message $m_{t,i} = (\vec{G}_{t,i}, \mathbf{w}_{t,i})$:
  - Decide the status of $m_{t,i}$: $S_{t,i} \sim$ Bernoulli(λ_t)
  - If ($S_{t,i} = 0$)
    - Select a region: $R_{t,i} \sim$ Uniform($1/M$)
    - Generate a geo-tag: $\vec{G}_{t,i} \sim$ Uniform(f_0)
  - If (i=1$\wedge S_{t,1}$=1)$\vee$($i\geq$2$\wedge S_{t,i-1}$=0$\wedge S_{t,i}$=1)
    - Select a region: $R_{t,i} \sim$ Categorical($\vec{\delta}_0$)
    - Generate a geo-tag: $\vec{G}_{t,i} \sim N(\mu_{R_{t,i}}, \Sigma_{R_{t,i}})$
  - Else (i.e., $i\geq$2 $\wedge$ $S_{t,i-1} = 1 \wedge S_{t,i}$ =1)
    - Select a region: $R_{t,i} \sim$ Categorical($\vec{\delta}_{R_{t,i-1},Z_{t,i-1}}$)
    - Generate a geo-tag: $\vec{G}_{t,i} \sim N(\mu_{R_{t,i}}, \Sigma_{R_{t,i}})$
  - Select a topic: $Z_{t,i} \sim$ Categorical($\vec{\theta}_{R_{t,i}}$)
  - Generate a message: $\mathbf{w}_{t,i} \sim$ Multinomial($\vec{\phi}_{Z_{t,i}}$)

Figure 5: The generative process of our model

4.2 Likelihood

Let Ω be the set of posterior distributions and model parameters that we have to estimate. Given Ω, let $Pr_\Omega(s)$ represent the probability that a sequence $s{=}\langle m_{1:N}\rangle$ of geo-tagged messages is generated. By introducing latent variables $R_{1:N}$, $S_{1:N}$ and $Z_{1:N}$, we can derive the probability $Pr_\Omega(s)$ based on our model as follows:

$$Pr_\Omega(\langle m_{1:N}\rangle) = \sum_{R_{1:N}, S_{1:N}, Z_{1:N}} \lambda_{S_1} \cdot Pr(R_1|S_1) \cdot \theta_{R_1,Z_1} \cdot Pr(m_1|R_1,Z_1) \cdot \prod_{i=2}^{N} \lambda_{S_i} \cdot Pr(R_i|S_i,S_{i-1},R_{i-1},Z_{i-1}) \cdot \theta_{R_i,Z_i} \cdot Pr(m_i|R_i,Z_i)$$

where $Pr(m_i|R_i,Z_i)$ represents the probability that a geo-tagged message $m_i{=}(\vec{G}_i, \mathbf{w}_i)$ is generated given latent semantic region R_i and topic Z_i, which is

$$Pr(m_i|R_i,Z_i) = f_{R_i}(\vec{G}_i) \cdot \prod_{j=1}^{N_i} \phi_{Z_i,w_j}, \tag{2}$$

and $Pr(R_i|S_i,S_{i-1},R_{i-1},Z_{i-1})$ (or $Pr(R_i|S_i)$ with $i{=}1$) is the probability to select a region R_i given S_{i-1}, S_i, R_{i-1} and Z_{i-1}, which is formulated as

$$Pr(R_i|S_i,S_{i-1},R_{i-1},Z_{i-1}) = \tag{3}$$
$$\begin{cases} 1/M & \text{if } S_i{=}0, \\ \bar{\delta}_{R_i} & \text{if } (i{=}1 \text{ and } S_i{=}1) \text{ or } (i{\geq}2 \text{ and } S_{i-1}{=}0), \\ \delta_{R_{i-1},Z_{i-1},R_i} & \text{if } i{\geq}2 \text{ and } S_i{=}1 \text{ and } S_{i-1}{=}1. \end{cases}$$

Then, the likelihood $\mathbb{L}$ can be directly formulated as

$$\mathbb{L} = \prod_{k=1}^{K} \int_{\vec{\phi}_k} Dir(\vec{\phi}_k;\vec{\beta}) \prod_{r=1}^{M} \int_{\vec{\theta}_r} Dir(\vec{\theta}_r;\vec{\alpha}) \prod_{t=1}^{t} \int_{\vec{\lambda}_t} Beta(\vec{\lambda}_t;\vec{\gamma}) Pr_\Omega(s_t) d\vec{\lambda}_t d\vec{\theta}_r \vec{\phi}_k \tag{4}$$

where Dir and $Beta$ denote the Dirichlet and Beta distribution respectively.

4.3 Variational EM Algorithm

In order to estimate the most likely posterior distributions and model parameters with a given collection of trajectories, we utilize the variational EM algorithm[9], which is one of the most popular methods for maximum likelihood estimation. To apply the variational EM algorithm, we perform the mean field approximation [2]. We summarized the approximate variables for the posterior distributions in Figure 4. Then, we can derive a lower bound of log-likelihood $\mathbb{F}$ by utilizing Jensen's inequality[15]. We next derive the update equations of the EM step for every model parameter and approximate variable to maximize $\mathbb{F}$. Since they include long equations, we provide the lower bound $\mathbb{F}$ and the update equations in Appendix not to break the flow of reading.

4.4 Finding the Most Likely Sequence

To find the significant transition patterns based on the estimated model parameters, we now present a dynamic programming algorithm which computes the most likely sequence of latent semantic regions for a given trajectory. This is similar to the Viterbi algorithm for the HMM model[12], which finds the most likely sequence of hidden states that results in the sequence of observed events.

Dynamic programming: Given a sequence s_t with n geo-tagged messages, we use the following notations in our dynamic programming formulation.

- $s_t[i]$: the subsequence of s_t which starts at the first message and ends at the i-th message of s_t.
- $\bar{\pi}[i]$: the maximum probability to generate $s_t[i]$ when $m_{t,i}$ is submitted without any local context (i.e., $S_{t,i}{=}0$).
- $\pi[i,r,k]$: the maximum probability to create $s_t[i]$ when $m_{t,i}$ has the local context, its latent semantic region is r and the latent topic is k (i.e., $S_{t,i}{=}1 \wedge R_{t,i}{=}r \wedge Z_{t,i}{=}k$).
- $\Pi[i]$: the maximum probability to generate $s_t[i]$ which is computed as $\max\{\bar{\pi}[i], \max_{1\leq r\leq M, 1\leq k\leq K} \pi[i,r,k]\}$.

Computing $\bar{\pi}[i]$ and $\pi[i,r,k]$: Let $Pr(m_{t,i}|R_{t,i}{=}r, Z_{t,i}{=}k)$ denote the probability of posting $m_{t,i}$ in the latent semantic region r with topic k as defined in Equation (2). We first consider the case when $m_{t,i}$ is not concerned with a local context (i.e., $S_{t,i}{=}0$). Regardless of which semantic regions and topics were involved to generate its preceding message $m_{t,i-1}$, the maximum probability $\bar{\pi}[i]$ is simply determined by $R_{t,i}$ and $Z_{t,i}$ both of which maximize $Pr(m_{t,i}|R_{t,i}{=}r, Z_{t,i}{=}k)$. Thus, $\bar{\pi}[i]$ is computed as

$$\bar{\pi}[i] = \Pi[i-1] \cdot \max_{1\leq r\leq M, 1\leq k\leq K} \lambda_0 \frac{1}{M} \theta_{r,k} Pr(m_{t,i}|R_{t,i}=r, Z_{t,i}=k).$$

Next, suppose that $m_{t,i}$ is generated in a local context (i.e., $S_{t,i}{=}1$). We need to consider two cases to select the most likely region $R_{t,i}$: (1) when the preceding message $m_{t,i-1}$ is not related to the local context (i.e., $S_{t,i-1}{=}0$) and (2) when it is (i.e., $S_{t,i-1}{=}1$). For case (1), we can compute the maximum probability $\pi[i,r,k]$ to generate $s_t[i]$ regardless of $R_{t,i-1}$ and $Z_{t,i-1}$ since we choose $R_{t,i}$ independent to its previous message in our model. For case (2), since $R_{t,i}=r$ is selected depending on $R_{t,i-1}$ and $Z_{t,i-1}$, we compute $\pi[i,r,k]$ by enumerating the maximum probability to generate $s_t[i]$ with every pair of $R_{t,i-1}$ and $Z_{t,i-1}$. Thus, the recursive solution to compute $\pi[i,r,k]$ is

$$\pi[i,r,k] = \max \left\{ \begin{array}{l} \Pi[i-1] \cdot \lambda_1 \cdot \delta_{0,r} \cdot \theta_{r,k} \cdot Pr(m_{t,i}|R_{t,i}=r, Z_{t,i}=k), \\ \qquad\qquad \text{Case (1)} \\ \max_{1\leq r'\leq M, 1\leq k'\leq K}\{\pi[i-1,r',k'] \cdot \lambda_1 \cdot \delta_{r',k',r} \cdot \theta_{r,k} \\ \cdot Pr(m_{t,i}|R_{t,i}=r, Z_{t,i}=k)\} \\ \qquad\qquad \text{Case (2)} \end{array} \right\}$$

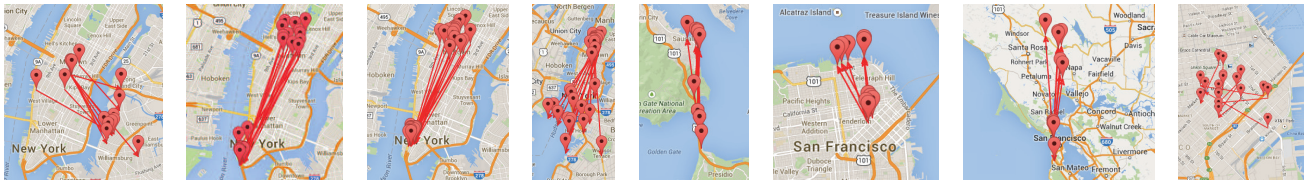
Note that since $\max_{1\leq r'\leq M, 1\leq k'\leq K}\{\ \pi[i-1,r',k'] \cdot \delta_{r',k',r}\}$ for case (2) can be calculated without depending on k, we can compute the part only once for each region r and use the maximum value for every k when we compute $\pi[i,r,k]$ for every pair of r and k. We refer to this dynamic programming algorithm as *TOPTRAC-MLS*.

Time complexity: Let n be the length of the sequence s. For each of the i-th position and each region r, since it takes $O(M{\cdot}K{+}K)$ time to compute $\pi[i,r,k]$ for all $k{=}1,\ldots,K$, the time complexity of *TOPTRAC-MLS* becomes $O(n{\cdot}M^2{\cdot}K)$.

4.5 Finding Frequent Transition Patterns

We next find frequent transition patterns based on the most likely sequences and then, for each transition pattern, we select the top-k transition snippets which best represent the pattern.

Let $\bar{s}_t{=}\langle(s_{t,1},r_{t,1},z_{t,1}),\ldots,(s_{t,N_t},r_{t,N_t},z_{t,N_t})\rangle$ denote the most likely sequence of latent variables computed by *TOPTRAC-MLS* for each sequence s_t in the trajectory collection $\mathbb{C}$, where $(s_{t,i},r_{t,i},z_{t,i})$ represents the values of $S_{t,i}$, $R_{t,i}$ and $Z_{t,i}$ respectively. We use $S_{ML}{=}\{\bar{s}_1,\ldots,\bar{s}_T\}$ to denote the set of the most likely sequence of every trajectory in $\mathbb{C}$. We define *transition pattern* and *transition snippet* as follows:



(a) Brooklyn (b) Battery Park (c) Ctr. Memorial (d) LGTA (e) Sausalito (f) Chinatown (g) Winery (h) NAIVE

Figure 6: Transition patterns in NYC and SANF

DEFINITION 4.1.: *Let s_t and $\bar{s}_t$ be a trajectory and its most likely sequence respectively. Given a maximum time interval Δ, a transition pattern $\langle(r_1, z_1), (r_2, z_2)\rangle$ $(r_1 \neq r_2)$ is said to be* supported *by s_t if and only if*

(1) $\bar{s}_t$ includes a subsequence $\langle(s_{t,i}, r_{t,i}, z_{t,i}), \ldots, (s_{t,j}, r_{t,j}, z_{t,j})\rangle$ which starts with $(1, r_1, z_1)$ and ends with $(1, r_2, z_2)$ (i.e., $(s_{t,i} = 1 \wedge r_{t,i} = r_1 \wedge z_{t,i} = z_1) \wedge (s_{t,j} = 1 \wedge r_{t,j} = r_2 \wedge z_{t,j} = z_2)$)

(2) $s_i.timestamp + \Delta \geq s_j.timestamp$

Furthermore, the transition snippet *of the pattern is defined as the pairs of geo-tagged messages $\langle m_{t,i}, m_{t,j}\rangle$ in s_t that supports the pattern.*

Given S_{ML} and the minimum support τ, we define the *frequent transition pattern* as a subsequence of two topical semantic regions $\langle(r_1, z_1), (r_2, z_2)\rangle$ which is supported by at least τ trajectories in $\mathbb{C}$.

EXAMPLE 4.2.: *Suppose that we have obtained two most likely sequences from $\mathbb{C}$: $\bar{s}_1 = \langle(0, 1, 1), (1, 1, 2), (1, 2, 1)\rangle$ and $\bar{s}_2 = \langle(1, 1, 2), (0, 2, 1), (1, 2, 1)\rangle$. Assume that the time interval Δ is 6 hours and every geo-tagged message is submitted within 6 hours. Let $\tau = 2$. Then, $\langle(1, 2), (2, 1)\rangle$ is a frequent transition pattern since both s_1 and s_2 support the pattern.*∎

Among all transition snippets $\langle m_{t,i}, m_{t,j}\rangle$ of the frequent pattern $\langle(r_1, z_1), (r_2, z_2)\rangle$ appearing in any $s_t \in \mathbb{C}$, the *top-k transition snippets* are the k transition snippets which have the k largest probabilities of

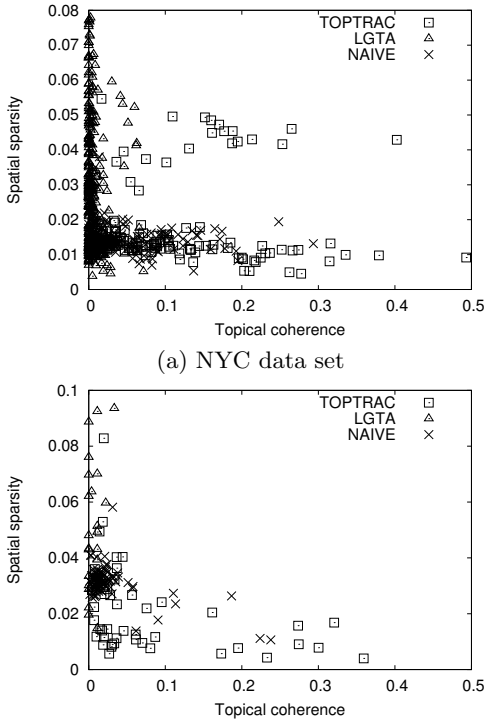
$$\delta_{r_1,z_1,r_2} Pr(m_{t,i}|R_{t,i} = r_1, Z_{t,i} = z_1) Pr(m_{t,j}|R_{t,j} = r_2, Z_{t,j} = z_2),$$

which represents the probability that a user posts the geo-tagged message $m_{t,i}$ with respect to the topic z_1 in the semantic region r_1 and next transports himself to r_2 where he submits $m_{t,j}$ with topic z_2.

5. EXPERIMENTS

We empirically evaluated the performance of our proposed algorithms. All experiments reported in this section were performed on the machines with Intel(R) Core(TM)2 Duo CPU 2.66GHz and 8GB of main memory. All algorithms were implemented using Javac Compiler of version 1.7. For our experiments, we implemented the following algorithms.

- **TOPTRAC**: This is the implementation of our trajectory pattern mining algorithm proposed in this paper.
- **LGTA**: It denotes the extended *LGTA* algorithm[17] to find trajectory patterns. In this algorithm, we first run the inference algorithm in [17] and find frequent trajectory patterns similar to our algorithm in Section 4.4. We present the details of the algorithm in Appendix A.



(a) NYC data set

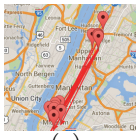
(b) SAF data set

Figure 8: Distributions of transition patterns

- **NAIVE**: We implemented a naive trajectory pattern mining algorithm which first groups messages geographically by GPS locations using EM clustering[1] and then clusters the messages topically in each group with LDA[2]. In each geographical cluster r, we select the most likely latent topic for a message $m_{t,i}$ by using $\arg\max_{k=1,..,K} \theta_{r,k} \prod_{j=1}^{N_{t,i}} \phi_{k,w_{t,i,j}}$ where $\theta_{r,k}$ and $\phi_{k,w}$ are the model parameters of LDA. Then, transition patterns and their top-k snippets are computed similarly to *TOPTRAC*.

5.1 Data Sets

For real-life data sets, we collected geo-tagged messages using the Streaming API of Twitter[14] from May, 25th to May, 31st in 2014. We used two data sets of geo-tagged messages collected in two cities, which are New York city and San Francisco, USA and we will denote them as *NYC* and *SANF* respectively. The data set *NYC* includes 9,070 trajectories with 266,808 geo-tagged messages. In *SANF*, there are 809 trajectories with 19,664 messages. For both data sets, we eliminated stop-words which appear less than 5 times or occur in more than 10% of all messages and removed the messages including some words which obviously have no local context such as insults.



(a)

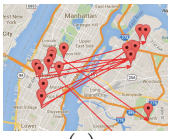
(1) @ Yankee Stadium http://t.co/*
(2) This was for dinner. #NYC #bestrestaurant #foodporn @ The Halal Guys http://t.co/*
(1) I'm at Yankee Stadium - @mlb (Bronx, NY) http://t.co/*
(2) I'm at Wa Bar + Kitchen - @wabarnyc (New York, NY) http://t.co/*
(1) We are here for the gamew (@ Yankee Stadium Gate 8 w/ 7 others) http://t.co/*
(2) I'm at O'Donoghues Bar & Restaurant - @odonoghuests (New York, NY) w/ 2 others http://t.co/*

(b) From 'Yankee Stadium' to Downtown for 'Dinner'

(c)

(1) Morning @ Central Park 5av and 59th Street NYC http://t.co/*
(2) Just posted a photo @ Egg Restaurant http://t.co/*
(1) I'm at The Biergarten at The @StandardHotels (New York, NY) w/ 17 others http://t.co/*
(2) Brunching (at @Allswell_nyc) http://t.co/*
(1) #nyc icon. @ Empire State Building http://t.co/*
(2) Brunch time. #nyc #brooklyn @ Juliette Restaurant http://t.co/*

(d) From Downtown to Brooklyn for 'Brunch'



(e)

(1) I'm at Apple Store (New York, NY) w/ 7 others http://t.co/*
(2) I'm at @DunkinDonuts (Maspeth, NY) https://t.co/*
(1) I'm at Chipotle Mexican Grill (New York, NY) w/ 4 others http://t.co/*
(2) I'm at Brooklyn Bagel & Coffee Company (Astoria, NY) w/ 2 others http://t.co/*
(1) I'm at Union Square Park - @nycparks (New York, NY) w/ 31 others http://t.co/*
(2) I'm at New York City Bagel & Coffee House - @nycbch (Astoria, NY) http://t.co/*

(f) From Downtown to Brooklyn for 'Bagel'

Figure 7: Interesting transition patterns in NYC

5.2 Quality of Transition Patterns

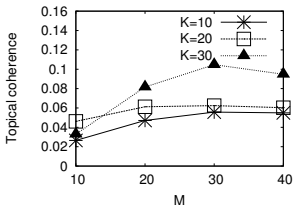
We conducted our experiments with varying the number of latent semantic regions M, the number of latent topics K and the minimum support τ. The default values of these parameters for each data set were: M=30, K=30, τ=100 for *NYC*, and M=20, K=20, τ=10 for *SANF*. The time interval Δ between two regions in a transition was set to 6 hours and the default value of k for the top-k representative transition snippets was 15. Furthermore, the hyper-parameters $\vec{\alpha}$, $\vec{\beta}$ and $\vec{\gamma}$ were updated between the iterations of the EM step using Newton method as LDA[2] does.

Quality measures: We computed the *spatial sparsity*, *topical coherence* and *topical anti-diversity* for evaluating the quality of each transition pattern which have been utilized to measure the quality of clusters[13]. Given a pattern with top-k transition snippets $\{\langle(\vec{G}_i^{from}, \mathbf{w}_i^{from}), (\vec{G}_i^{to}, \mathbf{w}_i^{to})\rangle\}_{i=1,..,k}$, the spatial sparsity is the average distance between every pair of geo-tags in each set of $\vec{G}_i^{from}$s and $\vec{G}_i^{to}$s, which is

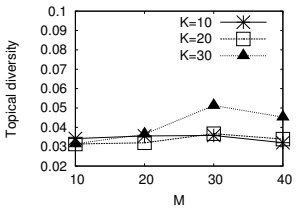
$$\frac{1}{2}\frac{2}{k(k-1)}\left\{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} d(\vec{G}_i^{from}, \vec{G}_j^{from})) + \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} d(\vec{G}_i^{to}, \vec{G}_j^{to}))\right\}$$

where $d(\vec{G}_1, \vec{G}_2)$ denotes the Euclidean distance between 2-dimensional vectors. Furthermore, the topical coherence is the average Jaccard similarity between every pair of messages in each set of $\mathbf{w}_i^{from}$s and $\mathbf{w}_i^{to}$s, which can be computed similarly to the spatial sparsity by substituting $d(\vec{G}_1, \vec{G}_2)$ with the Jaccard similarity $sim(\mathbf{w}_1, \mathbf{w}_2)$ in the above equation. The *topical anti-diversity* is the similarity between the different semantic region shown in the transition patterns. Given a pair (r, z) of the latent semantic region and topic, let $A_{(r,z)}$ denote the set of words in the messages chosen for the top-k transition snippets which are posted in r with topic z. The topical anti-diversity is defined as the average Jaccard similarity between $A_{(r,z)}$ and $A_{(r',z')}$ with every different (r, z) and (r', z') which have been appeared in the snippets.

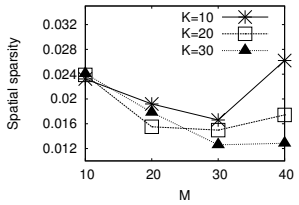
Illustrative Cases: We first present the interesting transition patterns found by *TOPTRAC* using real-life data which are selected among the patterns whose topical coherence are larger than 0.05. We show three transition patterns with their top-15 snippets found from *NYC* in Figures 6(a)-(c). The pattern shown in Figure 6(a) captures the movement of people who first visit 'Empire State Bldg.' and then transport themselves to 'Williamsburg waterfront pier' to see the Manhattan skyline. The representative snippets selected for this pattern are ('Sunset on the Empire', 'Just a photo @ Williamsburg Waterfront') and ('NYC at night from top of the Empire State Building', 'That Manhattan skyline').



(a) Topical coherence



(b) Topical anti-diversity



(c) Spatial sparsity

Figure 9: Quality of *TOPTRAC* varying M and K

In Figures 6(b)-(c), we can see that *TOPTRAC* splits the latent semantic regions for 'Battery park' and '9/11 Memorial center' desirably which are located close to each other, while the extended traditional method *LGTA* fails to split them as shown in Figure 6(d). This is because our model gives larger weights on the latent semantic regions which are visited by many people who moved in from another identical region with a coherent topic. Usually, tourists or people looking for popular restaurants, coffee shops or attractions create useful and interesting patterns.

To obtain finer granularities for *LGTA* and *NAIVE*, we can use larger number of regions in each method. However, as we increase the number of regions, we found that the transitions patterns obtained by these algorithms are mostly trivial and boring. In Figure 6(h), we present a typical transition pattern discovered by *NAIVE* when the number of regions was 70 for the data set *SANF*. It is a trivial transition

TOPTRAC					LGTA					NAIVE				
M	Q1	Q2	Q3	Q4	M	Q1	Q2	Q3	Q4	M	Q1	Q2	Q3	Q4
20	0.0816	0.0178	0.0576	0.0364	70	0.0048	0.023	0.0602	0.0068	70	0.042	0.0142	0.0326	0.0133
30	0.1047	0.0126	0.0459	0.0512	100	0.006	0.0301	0.0595	0.0067	100	0.061	0.0138	0.038	0.022
40	0.102	0.0124	0.0463	0.0548	150	0.0061	0.025	0.05	0.008	150	0.057	0.013	0.0293	0.0169

(a) NYC data set

TOPTRAC					LGTA					NAIVE				
M	Q1	Q2	Q3	Q4	M	Q1	Q2	Q3	Q4	M	Q1	Q2	Q3	Q4
10	0.0867	0.0293	0.0941	0.0396	30	0.0272	0.7340	0.8454	0.0134	50	0.0072	0.6274	0.678	0.015
20	0.0772	0.0196	0.0828	0.0464	40	0.0277	0.4553	0.4928	0.0125	70	0.0414	0.0261	0.0707	0.0487
30	0.0915	0.0172	0.1018	0.0501	50	0.0072	0.3274	0.6780	0.015	100	0.0256	0.0311	0.0581	0.0522

(b) SANF data set

Figure 10: Quality of clusters (Q1:topical coherence, Q2: spatial sparsity, Q3: distance, Q4: topical anti-diversity)

pattern shown in a small area with messages such as 'You bet!' or 'Lunch time'. In contrast, *TOPTRAC* found the interesting patterns including not only small latent semantic regions such as 'Sausalito' in Figure 6(e) or 'Fisherswarf' and 'Chinatown' in Figure 6(f) but also the regions with much larger area such as 'Winery' as shown in Figure 6(g).

Furthermore, we present more illustrative cases found in *NYC* in Figure 7 with their transition snippets. These cases show that *TOPTRAC* can discover interesting topical transition patterns with good representative snippets.

Comparative study: By varying the number of regions M, we evaluated the performance of the implemented algorithms in terms of spatial sparsity, topical coherence and topical anti-diversity, and presented the result in Figure 10. We set K and τ to the default values for each data set. The range of M in each algorithm is selected as the values with which each one achieves the best performance. The result confirms that *TOPTRAC* outperforms the other algorithms especially with respect to topical coherence.

NAIVE obtained the next largest topical coherence. However, this is because if M is large, it finds transition patterns appearing in a very small area created by active users who post very personal and daily talks as we discussed with the illustrative case in Figure 6(h). As evidence, we can find that the average distance (Q3) between starting and ending locations of patterns becomes smaller with growing M.

In addition, *LGTA* was even worse than *NAIVE* algorithm in terms of both quality measures: topical coherence and spatial sparsity. This is because *LGTA* is sensitive to noisy messages (i.e., daily talks without local contexts) and sparsely distributed geo-tags since *LGTA* tends to find large semantic regions to capture the common topics among the noisy messages.

Distribution of patterns: With *NYC* and *SANF*, we plotted the transition patterns in Figure 8 where x-axis and y-axis are the topical coherence and spatial sparsity respectively. The graphs show that *TOPTRAC* identifies transition patterns with high topical coherence and low spatial sparsity well while *LGTA* and *NAIVE* fail to find any transition pattern with high topical coherence. Most of the patterns found by *LGTA* and *NAIVE* also obtained small spatial sparsity, simply because trivial patterns are captured in small areas by using a large number of regions as we discussed before.

Furthermore, the graphs show that *TOPTRAC* is able to find transition patterns which not only indicate strong topical coherence but also involve large semantic regions well, such as the transition to 'Winery' shown in Figure 6(g).

Varying M and K for TOPTRAC: With varying the numbers of regions M and topics K, we plotted the quality of *TOPTRAC* in Figure 9 for *NYC* data set. As we increase M, the topic coherence grows gradually with every range of K as shown in Figure 9(a) since *TOPTRAC* can generate more semantic regions with large M. The spatial sparsity is decreased with varying M from 10 to 30 but rises when M is larger than 30 in in Figure 9(c). This is because the messages posted without local context happen to be clustered in large latent semantic regions as M is large enough to group those noisy messages into semantic regions. We actually have confirmed that the latent semantic regions with a large determinant are produced when M is large.

Furthermore, the larger K we use, the higher topical coherence we obtain because we can split semantic regions with a finer granularity. However, if we set larger value of K than 30, a topic can be divided into several topics (i.e., a single topic can be represented with a mixture of more than a topic), and it results in that few patterns are identified by our *TOPTRAC-MLS* algorithm. Thus, the graphs confirms that *TOPTRAC* achieved the best performance with M=30 and K=30 which are the default setting for *NYC*. The default setting for *SANF* was also determined similarly.

6. CONCLUSION

In this paper, we proposed a trajectory pattern mining algorithm, called *TOPTRAC*, using our probabilistic model to capture the spatial and topical patterns of users, who transport themselves while posting geo-tagged messages in micro-blogging services. We developed an efficient inference algorithm for our model and also devised algorithms to find frequent transition patterns as well as the best representative snippets of each pattern. Our experiments using real-life data sets confirmed that our method not only discovers useful and interesting transition patterns, but also identifies important semantic regions with refined granularity.

Acknowledgment

This work was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1017362, IIS-1320617, and IIS-1354329, HDTRA1-10-1-0120, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), and MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC. It was also partially supported by the research fund of Hanyang University(HY-2014-N) and Aero-Science Fund of China-2013ZD51058.

7. REFERENCES

[1] J. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, 1998.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] Z. Chen, H. T. Shen, and X. Zhou. Discovering popular routes from trajectories. In *ICDE*, 2011.

[4] F. Giannotti, M. Nanni, D. Pedreschi, and F. Pinelli. Trajectory pattern analysis for urban traffic. In *IWCTS*, 2009.

[5] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *SIGKDD*, 2007.

[6] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999.

[7] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *WWW*, 2012.

[8] B. Hu and M. Ester. Spatial topic modeling in online social media for location recommendation. In *RecSys*, 2013.

[9] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[10] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura. Travel route recommendation using geotags in photo sharing sites. In *CIKM*, 2010.

[11] J.-G. Lee, J. Han, and K.-Y. Whang. Trajectory clustering: a partition-and-group framework. In *SIGMOD Conference*, 2007.

[12] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *PROCEEDINGS OF THE IEEE*, 1989.

[13] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *SIGKDD*, 2009.

[14] Twitter. Streaming api. http://dev.twitter.com/, 2014.

[15] Wikipedia. Jensen's inequality. http://en.wikipedia.org/wiki/Jensen's_inequality, 2013.

[16] Z. Yin, L. Cao, J. Han, J. Luo, and T. S. Huang. Diversified trajectory pattern ranking in geo-tagged social media. In *SDM*, 2011.

[17] Z. Yin, L. Cao, J. Han, C. Zhai, and T. S. Huang. Geographical topic discovery and comparison. In *WWW*, 2011.

[18] C. Zhang, J. Han, L. Shou, J. Lu, and T. L. Porta. Splitter: Mining finegrained sequential patterns in semantic trajectories. In *VLDB*, 2014.

[19] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *WWW*, 2009.

APPENDIX

A. AN EXTENSION OF LGTA

In [17], Yin et al. proposed an algorithm, called *LGTA*, to discover latent semantic regions using the GPS location and user-generated tags submitted to the photo-sharing services such as Flickr. Without considering any dependency between messages in sequences, *LGTA* discovers the region in which the messages are posted with the same topic preference. Since their model focuses on finding latent semantic regions only, *LGTA* does not provide a method to find transition patterns. However, by utilizing the simple Markov model as proposed in [10], we can extend *LGTA* to handle our transition pattern mining problem.

LGTA assumes that when the number of latent semantic regions is M, the messages posted in the same latent semantic region are generated by following the same mixture of topic distributions, where the region is also determined probabilistically. Let $r_{t,i}$ denote the latent semantic region of the message $m_{t,i}$=($\vec{G}_{t,i}$, $\mathbf{w}_{t,i}$). For each geo-tagged message $m_{t,i}$, words in the message $\mathbf{w}_{t,i}$ are selected repeated similarly to PLSI[6] or LDA[2] model and the geo-tag $\vec{G}_{t,i}$ is chosen by the Gaussian distribution of region $r_{t,i}$.

Extended LGTA: Using the model parameters estimated by LGTA, we find transition patterns and their snippets as follows: After running the inference algorithm in [17], for each geo-tagged message $m_{t,i}$, we can compute the most likely semantic region r which maximizes the likelihood $Pr(m_{t,i}|R_{t,i}=r)$, which is the probability to submit $m_{t,i}$ in the region r. Let S_{ML} be the set of the sequences of the most likely latent semantic region $r_{t,i}$ for each message $m_{t,i}$ in $\mathcal{C}$. Given S_{ML} and the minimum support τ, *transition patterns* are every subsequence with length 2, $\langle r_1, r_2\rangle$ with $r_1 \neq r_2$ and $1 \leq r_1, r_2 \leq M$, which occurs at least τ times in S_{ML} within a given time interval Δ.

We next compute the top-k transition snippets for each transition pattern. Let $Pr(r_1 \rightarrow r_2)$ denote the conditional probability that a user posts one message in r_1 and then submits another in r_2 within the interval of Δ. Then, $Pr(r_1 \rightarrow r_2)$ is calculated as

$$Pr(r_1 \rightarrow r_2) = \frac{N(r_1, r_2)}{N(r_1)}$$

where $N(r_1)$ and $N(r_1, r_2)$ denote the frequencies of the latent semantic region r_1 and the subsequence $\langle r_1, r_2\rangle$ appearing within Δ in S_{ML} respectively. Then, the *top-k transition snippets* of each transition pattern $\langle r_1, r_2\rangle$ are the k pairs of geo-tagged messages $\langle m_{t,i}, m_{t,i+1}\rangle$ in $\mathcal{C}$ whose probabilities of posting $m_{t,i}$ in r_1 and moving to r_2 to submit $m_{t,i+1}$:

$$Pr(r_1 \rightarrow r_2) \cdot Pr(m_{t,i}|R_{t,i}=r_1) \cdot Pr(m_{t,i+1}|R_{t,i+1}=r_2)$$

are the k largest.

B. THE LOWER BOUND OF LOG-LIKELIHOOD

By introducing the approximate parameters $\vec{a}_t$, $\vec{b}_k$, $\vec{c}_t$, $\vec{\sigma}_{t,i}$, $\vec{\rho}_{t,i}$ and $\vec{\zeta}_{t,i}$ as summarized in Figure 4, we obtain the lower bound of log-likelihood $\mathbb{F}$ as follows:

$$
\begin{aligned}
\mathbb{F} = & \\
& \sum_{t=1}^{N}\sum_{i=1}^{N_t}\sum_{s=0}^{1}\{\sigma_{t,i,s}(\Psi(c_{t,s}) - \Psi(c_{t,\bullet})) - \sigma_{t,i,s}\log\sigma_{t,i,s}\} \\
& + \sum_{t=1}^{N}\sum_{i=1}^{N_t}\sigma_{t,i,0}\log 1/M + \sum_{t=1}^{N}\sum_{r=1}^{M}\sigma_{t,1,1}\rho_{t,1,r}\log\bar{\delta}_r \\
& + \sum_{t=1}^{N}\sum_{i=2}^{N_t}\sum_{r=1}^{M}\sigma_{t,i-1,0}\sigma_{t,i,1}\rho_{t,i,r} log\bar{\delta}_r \\
& + \sum_{t=1}^{N}\sum_{i=2}^{N_t}\sum_{r'=1}^{M}\sum_{k=1}^{K}\sum_{r=1}^{M}\sigma_{t,i-1,1}\sigma_{t,i,1}\rho_{t,i-1,r'}\rho_{t,i,r}\zeta_{t,i-1,k}\log\delta_{r',k,r} \\
& + \sum_{t=1}^{N}\sum_{i=1}^{N_t}\sum_{r=1}^{M}\sigma_{t,i,1}\rho_{t,i,r}\log f_r(\vec{G}_{t,i}) - \sum_{t=1}^{N}\sum_{i=1}^{N_t}\sum_{r=1}^{M}\rho_{t,i,r}\log\rho_{t,i,r} \\
& + \sum_{t=1}^{N}\sum_{i=1}^{N_t}\sum_{k=1}^{K}\left\{\sum_{r=1}^{M}\rho_{t,i,r}\zeta_{t,i,k}(\Psi(a_{r,k}) - \Psi(a_{r,\bullet})) - \zeta_{t,i,k}\log\zeta_{t,i,k}\right\} \\
& + \sum_{t=1}^{N}\sum_{i=1}^{N_t}\sum_{j=1}^{N_{t,i}}\sum_{k=1}^{K}\zeta_{t,i,k}(\Psi(b_{k,w_{t,i,j}}) - \Psi(b_{k,\bullet}))
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{r=1}^{M}\left\{\log\Gamma(\alpha_\bullet)-\sum_{k=1}^{K}\log\Gamma(\alpha_k)+\sum_{k=1}^{K}(\alpha_k-1)(\Psi(a_{r,k})-\Psi(a_{r,\bullet}))\right\}\\
&-\sum_{r=1}^{M}\left\{\log\Gamma(a_{r,\bullet})-\sum_{k=1}^{K}\log\Gamma(a_{r,k})+\sum_{k=1}^{K}(a_{r,k}-1)(\Psi(a_{r,k})-\Psi(a_{r,\bullet}))\right\}\\
&+\sum_{k=1}^{K}\left\{\log\Gamma(\beta_\bullet)-\sum_{v=1}^{V}\log\Gamma(\beta_v)+\sum_{v=1}^{V}(\beta_v-1)(\Psi(b_{k,v})-\Psi(b_{k,\bullet}))\right\}\\
&-\sum_{k=1}^{K}\left\{\log\Gamma(b_{k,\bullet})-\sum_{v=1}^{V}\log\Gamma(b_{k,v})+\sum_{v=1}^{V}(b_{k,v}-1)(\Psi(b_{k,v})-\Psi(b_{k,\bullet}))\right\}\\
&+\sum_{t=1}^{N}\left\{\log\Gamma(\gamma_\bullet)-\sum_{s=0}^{1}\log\Gamma(\gamma_s)+\sum_{s=0}^{1}(\gamma_s-1)(\Psi(c_{t,s})-\Psi(c_{t,\bullet}))\right\}\\
&-\sum_{t=1}^{N}\left\{\log\Gamma(c_{t,\bullet})-\sum_{s=0}^{1}\log\Gamma(c_{t,s})+\sum_{s=0}^{1}(c_{t,s}-1)(\Psi(c_{t,s})-\Psi(c_{t,\bullet}))\right\}
\end{aligned}
\tag{5}
$$

where $a_{r,\bullet}$, $b_{k,\bullet}$ and $c_{t,\bullet}$ represent $\sum_{k=1}^{K} a_{r,k}$, $\sum_{v=1}^{V} b_{k,v}$ and $\sum_{s=0}^{1} c_{t,s}$ respectively.

C. DERIVATION OF EM STEPS

In the variational EM Steps, we calculate the variational parameters maximizing the lower bound $\mathbb{F}$ in Equation (5) repeatedly. In this section, we derive the update equations for the variational parameters $a_{t,k}$, $b_{k,v}$, $c_{t,s}$, $\sigma_{t,i,s}$, $\rho_{t,i,r}$ and $\zeta_{t,i,k}$ by using the method of Lagrange multipliers.

Update equations of $a_{r,k}$, $b_{k,v}$ and $c_{t,s}$: We first compute the derivative $\mathbb{F}$ with respect to $a_{r,k}$ as follows.

$$
\begin{aligned}
&\frac{\partial\mathbb{F}}{\partial a_{r,k}}\\
&=\Psi'(a_{r,k})\left(\alpha_{a,k}+\sum_{t=1}^{N}\sum_{i=1}^{N_t}\rho_{t,i,r}\zeta t,i,k-a_{r,k}\right)\\
&+\Psi'(a_{r,\bullet})\sum_{k'=1}^{K}\left(\alpha_{a,k'}+\sum_{t=1}^{N}\sum_{i=1}^{N_t}\rho_{t,i,r}\zeta t,i,k'-a_{r,k'}\right)=0
\end{aligned}
$$

Then, the solution of $a_{r,k}$ which always satisfies the above equation is

$$
a_{r,k}=\alpha_k+\sum_{t=1}^{T}\sum_{i=1}^{N_t}\rho_{t,i,r}\zeta_{t,i,k} \tag{6}
$$

In a similar way, the update equations of $b_{k,v}$ and $c_{t,s}$ can be derived as follows:

$$
b_{k,v}=\beta_v+\sum_{t=1}^{T}\sum_{i=1}^{N_t}\sum_{j=1}^{N_{t,i}}I_{(w_{t,i,j}=v)}\zeta_{t,i,k} \tag{7}
$$

$$
c_{t,s}=\gamma_s+\sum_{i=1}^{N_{t,i}}\sigma_{t,i,s} \tag{8}
$$

where $I_{condition}$ denotes the indicator function which gives 1 if the *condition* holds and 0 otherwise.

Update of $\sigma_{t,i,s}$: Similarly, we next derive the update equation of $\sigma_{t,i,s}$, which is a variational parameter for the distribution $q(R_{t,i})$ as shown in Figure 4, by differentiating $\mathbb{F}$ with $\sigma_{t,i,s}$ as shown below:

$$
\begin{aligned}
\sigma_{t,i,0}&\propto\exp\left\{\Psi(c_{i,0})+\log\frac{1}{M}+\sum_{r=1}^{M}\sigma_{t,i+1,1}\rho_{t,i+1,r}\log\delta_{0,r}+\log f_0\right\}\\
\sigma_{t,i,1}&\propto\exp\left\{\Psi(c_{i,1})+\sum_{r=1}^{M}(\sigma_{t,i-1,0})^{I_{(i>1)}}\rho_{t,i,r}\log\delta_{0,r}\right.\\
&+\sum_{r=1}^{M}\sum_{k=1}^{K}\sum_{r'=1}^{M}(I_{(i>1)}\sigma_{t,i-1,1}\rho_{t,i-1,r}\zeta_{t,i-1,k}\rho_{t,i,r'}\\
&\left.+\sigma_{t,i+1,1}\rho_{t,i,r}\zeta_{t,i,k}\rho_{t,i+1,r'})\log\delta_{r,k,r'}+\sum_{r=1}^{M}\rho_{t,i,r}\log f_r(\vec{G}_{t,i})\right\}
\end{aligned}
\tag{9}
$$

Note that $\sigma_{t,i,0}+\sigma_{t,i,1}=1$ holds for every $t=1,\ldots,N$ and $i=1,\ldots,N_t$.

Update of $\rho_{t,i,r}$: The variational parameter $\rho_{t,i,r}$ of the distribution $q(R_{t,i})$ is computed as follows

$$
\begin{aligned}
\rho_{t,i,r}&\propto\exp\left\{(\sigma_{t,i-1,0})^{I_{i>1}}\sigma_{t,i,1}\log\delta_{0,r}\right.\\
&\sum_{r'=1}^{M}\sum_{k=1}^{K}(I_{i>1}\sigma_{t,i-1,1}\sigma_{t,i,1}\zeta_{t,i-1,k}\rho_{t,i-1,r'}\log\delta_{r',r}\\
&+\sigma_{t,i,1}\sigma_{t,i+1,1}\zeta_{t,i,k}\rho_{t,i+1,r'}\log\delta_{r,r'})\\
&\left.+\sum_{k=1}^{K}\zeta_{t,i,k}(\Psi(a_{r,k})-\Psi(a_{r,\bullet}))+\sigma_{t,i,1}\log f_r(\vec{G}_{t,i})\right\}
\end{aligned}
\tag{10}
$$

where $\sum_{r=1}^{M}\rho_{t,i,r}=1$ is satisfied for every $t=1,\ldots,N$ and $i=1,\ldots,N_t$.

Update of $\zeta_{t,i,r}$: We next take the derivative of $\mathbb{F}$ with regard to each of $\zeta_{t,i,k}$, which is a variational parameter for the distribution $q(Z_{t,i})$, and obtain the update equation as

$$
\begin{aligned}
\zeta_{t,i,k}&\propto\exp\left\{\sum_{r=1}^{M}\sum_{r'=1}^{M}\sigma_{t,i,1}\sigma_{t,i+1,1}\rho_{t,i,r}\rho_{t,i+1,r'}\log\delta_{r,k,r'}\right.\\
&\left.+\sum_{r=1}^{M}\rho_{t,i,r}\Psi(a_{r,k})+\sum_{j=1}^{N_{t,i}}(\Psi(b_{k,w_{t,i,j}})-\Psi(b_{k,\bullet}))\right\}
\end{aligned}
\tag{11}
$$

where $\zeta_{t,i,k}$ is normalized to satisfy $\sum_{k=1}^{K}\zeta_{t,i,k}=1$ for every $t=1,\ldots,N$ and $i=1,\ldots,N_t$.

Update of $\bar{\delta}_r$ and $\delta_{r,k,r'}$: We also obtain the update equations for $\bar{\delta}_r$ and $\delta_{r,k,r'}$ as

$$
\bar{\delta}_r\propto\sum_{t=1}^{T}\sum_{i=1}^{N_t}(\sigma_{t,i-1,0})^{I_{i>1}}\sigma_{t,i,1}\rho_{t,i,r} \tag{12}
$$

$$
\delta_{r,k,r'}\propto\sum_{t=1}^{T}\sum_{i=2}^{N_t}\sum_{k=1}^{K}\sigma_{t,i-1,1}\sigma_{t,i,1}\rho_{t,i-1,r}\rho_{t,i,r'}\zeta_{t,i-1,k} \tag{13}
$$

Note that $\sum_{r'=1}^{M}\bar{\delta}_{r'}=1$ and $\sum_{r'=1}^{M}\delta_{r,k,r'}=1$ for every $r=1,\ldots,M$ and $k=1,\ldots,K$.

Update of μ_r and Σ_r: We finally calculate μ_r and Σ_r, which is the center and covariance matrix of the r-th region defined in Equation (1), as

$$
\begin{aligned}
\vec{\mu}_r&=\frac{\sum_{t=1}^{T}\sum_{i=1}^{N_t}\sigma_{t,i,1}\rho_{t,i,r}\vec{G}_{t,i}}{\sum_{t=1}^{T}\sum_{i=1}^{N_t}\sigma_{t,i,1}\rho_{t,i,r}}\\
\Sigma_r&=\frac{\sum_{t=1}^{T}\sum_{i=1}^{N_t}\sigma_{t,i,1}\rho_{t,i,r}(\vec{G}_{t,i}-\vec{\mu}_r)(\vec{G}_{t,i}-\vec{\mu}_r)^\top}{\sum_{t=1}^{T}\sum_{i=1}^{N_t}\sigma_{t,i,1}\rho_{t,i,r}}
\end{aligned}
\tag{14}
$$