Minimizing User Involvement For Learning Human Mobility Patterns From Location Traces

Basma Alharbi, Abdulhakim Qahtan, Xiangliang Zhang

Computer, Electrical and Mathematical Sciences & Engineering Division
King Abdullah University of Science & Technology (KAUST)
Thuwal, 23955, Saudi Arabia

Abstract

Utilizing trajectories for modeling human mobility often involves extracting descriptive features for each individual, a procedure heavily based on experts' knowledge. In this work, our objective is to minimize human involvement and exploit the power of community in learning 'features' for individuals from their location traces. We propose a probabilistic graphical model that learns distribution of latent concepts, named *motifs*, from anonymized sequences of user locations. To handle variation in user activity level, our model learns motif distributions from sequence-level location co-occurrence of all users. To handle the big variation in location popularity, our model uses an asymmetric prior conditioned on per-sequence features. We evaluate the new representation in a link prediction task and compare our results to those of baseline approaches.

Introduction

With the increasing popularity of mobile devices and Location Based Social Networks (LBSNs), location traces for a large sample of the population become instantly available. The availability of such data allows for large-scale social analysis (Song et al. 2010), studies of human behavior (Gonzalez, Hidalgo, and Barabasi 2008) and urban analysis (Zheng, Zhang, and Yu 2015; Idé and Sugiyama 2011; Toole et al. 2012; Yuan, Zheng, and Xie 2012).

Although trajectory data convey indisputable patterns, employing them in basic applications usually involves layers of preprocessing and feature extraction. For example, the process followed to achieve a task such as demographic prediction often involves: 1) Trajectory preprocessing, which includes noise filtering, segmentation and stay point detection (Zheng 2015b). 2) Feature engineering, where trajectories are analyzed for extracting usable features, often based on rules initiated by experts. 3) Dimensionality reduction, for reducing the feature space. 4) Learning tasks are then completed in the reduced space. 5) Finally, these steps may be repeated until satisfactory results are obtained. Despite the evident advantage of this approach, minimizing human involvement is always preferable in knowledge execution
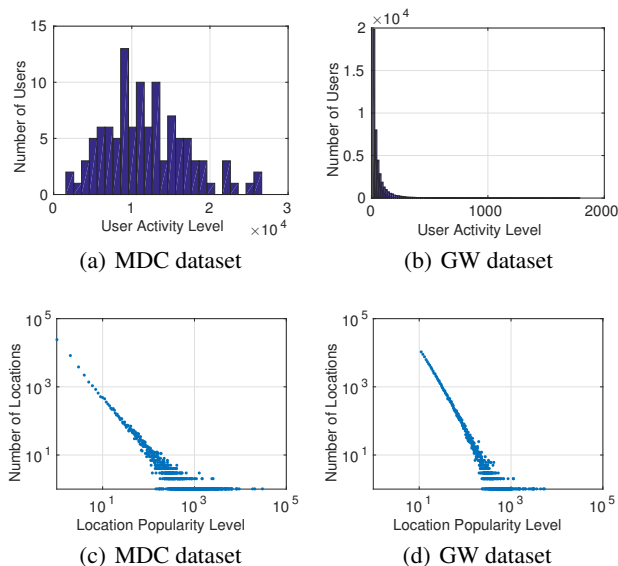
Figure 1: The distribution of user activity level and location popularity in MDC (a,c) and GW (b,d) dataset.

(Schumann and Lécué 2015). Rule-based feature engineering often ends up ignoring latent patterns which are important to the task under study but not obvious to the experts. More importantly, it is based on predefined knowledge of the data and the specific task under study, which restricts the use of extracted features to other applications.

Our objective in this work is to propose a representation learning model that infers latent patterns from user trajectories with minimum human intervention. The new representation can be used in various applications. Inspired by Probabilistic Graphical Models (PGM) in text mining, we propose a Human Mobility Representation model (HuMoR), which models trajectory data in an analogous manner to text corpora, where users are mapped to documents and locations are mapped to words. Similarly to modeling latent topics from documents using Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), our proposed model learns distribution of latent concepts, named *motifs*, from sequences of locations based on sequence-level location co-occurrence. Each user is then represented by a mixture of latent con-

cepts, i.e., a probability distribution over *motifs*. Comparing to building users' representation based on human efforts, the proposed approach is automated, efficient, and universally applicable.

HuMoR considers three key attributes of trajectory data:
1) User Activity Level: Users in a community have varying activity levels, ranging from very active users with dense trajectories to passive users with very sparse trajectories, as shown in Fig. 1(a) (the distribution of the number of visited locations for every user in the Mobile Data Challenge (MDC) Call-Detail Record (CDR) dataset (Laurila et al. 2012)). More challenges are brought by a large portion of inactive users, as observed in the Gowalla (GW) LBSN dataset (Cho, Myers, and Leskovec 2011) shown in Fig. 1(b), which has a skewed distribution over user activity levels. The difficulty of learning user representation here corresponds to the issue of learning from documents with different lengths, especially including many short text like tweets that LDA fails to learn (Tang et al. 2014). In our model, we introduce the notion of *sequences*, defined as segments extracted from users' trajectories. Extracted sequences have comparable lengths, and thus are much easier to handle than trajectories with varied activity levels.

2) Location Popularity Level: The popularity of locations, measured by the total number of visits, follows a power law distribution (shown in Fig. 1(c),1(d)). In text mining, there are also highly frequent words and rare words. Frequent words are usually eliminated to prevent them from dominating topics' distributions, as most of the frequent words are stop-words and their elimination has little impact on the context. However, frequent locations are not ignorable as they represent popular places. Consequently, our model considers frequent locations by adopting an asymmetric prior for the per-sequence motif distribution. This amendment allows some motifs to occur more often than others. As a result, popular locations can be coupled with such motifs, without dominating the remaining motif-location distributions. Asymmetric priors for LDA were discussed in (Wallach, Mimno, and McCallum 2009).

3) Side Features: Trajectory data contains informative side features which can be used to aid the learning process. This includes the timestamp of visited locations, often used to extract meaningful features such as home and work locations. Our model utilizes these features by conditioning a per-sequence prior over them, inspired by (Mimno and McCallum 2008). As our model imposes no restriction on the data source, incorporating cross-domain data classifies our model as an instance of data fusion approaches (Zheng 2015a).

Finally, HuMoR provides per-user motif distribution, which can be considered as a new representation for each user. We evaluate the new representation on social link prediction with two data sets, one of mobile records (MDC) and the other is extracted from an LBSN (GW). We compare our results against a number of baseline methods and show that our new representation has superior performance.

Related Work

Table 1 provides a summary of most relavent work. Recently, Kapicioglu et al. introduced Collaborative Place

Table 1: Summary of related work by the used method, the type of Location Description, the ability to handle Side Features and Variation in user activity and in location popularity.

Reference	Method	Location Description	Side Feature	Variation
Eagle et al.	PCA	Semantic	✗	-
Farrahi et al.	LDA	Semantic	✗	✗
Joseph et al.	LDA	Semantic	✗	✗
Phung et al.	LSDA	Semantic	✗	-
Kapicioglu et al.	CPM	Geographic	✗	-
Our Model	HuMoR	Anonymized	✓	✓

Models (CPM) to infer patterns from trajectories of geolocations, while other work uses semantically labeled locations. For example, Joseph, Tan, and Carley (2012) and Farrahi and Gatica-Perez (2011) employed LDA, Eagle and Pentland (2009) applied PCA instead of LDA, and Phung, Adams, and Venkatesh (2008) proposed *LSDA* (based on N-gram LDA), on semantic location traces.

Three factors distinguish our work from existing models,
1) **Learning from anonymized locations with minimum preprocessing of raw trajectories:** our model learns from anonymized locations (IDs of locations) as opposed to *semantically labeled locations*, which require human efforts, and *geographic locations*, which invade users' privacy.
2) **Handling side features and variations:** HuMoR enables adding per-sequence side features that enrich the learning process. In topic models like LDA, words in documents are assumed to be independent, i.e., documents are represented as a *bag-of-words*. Existing topic models that consider word order are mostly *n-gram* and/or Markovian models (Wallach 2006; Wang, McCallum, and Wei 2007; Gruber, Weiss, and Rosen-Zvi 2007) in which additional complexity is introduced due to the imposed dependency. Our model can consider order of visited locations without assuming a Markovian relation between location-IDs and/or motifs. This is achieved by introducing time range as a side feature. In addition, our model introduces an asymmetric prior and user *sequences* in the learning process to handle the variations in location popularity and user activity.
3) **Evaluation via real world application:** we assess the usefulness of the learned patterns in social link prediction. In related work, evaluation measures largely depend on listing the top locations per latent pattern, and visually assessing the similarity between them. We assert that a good representation is the one that captures the intrinsic characteristics of the high dimensional data without losing significant information. Thus, we evaluate our work on link prediction using two real datasets.

Human Mobility Representation Model
Preliminaries

Let $v \in \mathcal{V}$ be a user from the set of unique users in the data, and let $c \in \mathcal{C}$ be a location-ID from the set of unique location-IDs. User v's trajectory data, $\mathcal{R}(v)$, contains successive points of location-IDs visited over a period of time. Location-IDs have metadata (such as timestamps).

Definition 1 (User Social Graph) *Let $G = (\mathcal{V}, \mathcal{E})$ be a user social graph where nodes, $\mathcal{V}$, represent users and edges, $\mathcal{E}$, denote the social links between users.*

Definition 2 (Sequence) *Let $\boldsymbol{s} = (c_i, ..., c_{i+\mathcal{I}})$ be a sequence, where $\mathcal{I}$ is the length of the sequence. A sequence is defined as an ordered series of location-IDs, extracted from user v's trajectory $\mathcal{R}(v)$, in which all elements of the sequence share common side features. For example, if we consider time range as a side feature, $\boldsymbol{s}$ can be a sequence of location-IDs occurring in the morning. Given user v, $\boldsymbol{s}_v^{(j)}$ is user v's jth sequence such that $\boldsymbol{s}_v^{(j)} \subset \mathcal{R}(v)$. $\boldsymbol{\mathcal{S}}_v = \left\langle \boldsymbol{s}_v^{(j)} \right\rangle_{j=1}^{|\boldsymbol{\mathcal{S}}_v|}$ is user v's collection of sequences, and $\boldsymbol{\mathcal{S}} = \left\langle \boldsymbol{\mathcal{S}}_v \right\rangle_{v=1}^{|\mathcal{V}|}$ is the collection of sequences for all users.*

Problem Statement (User Representation) *Given users' trajectory data, $\mathcal{R}(.)$, learn compact and representative feature spaces, $\hat{\mathcal{R}}(.)$, for all users in the community, where the reduced feature space can be used as an alternative to the original one in applications such as link prediction.*

Model Description

To solve the stated problem, we exploit the existence of latent patterns in human mobility (Song et al. 2010), which we refer to as *motifs*. These motifs can be used to describe users in a compact feature space. As a result, we map the problem from learning users' representations, in general, to explicitly learning motifs from trajectory data. We propose to model motifs, in an analogous manner to modeling topics in text, using generative PGMs. We propose a Human Mobility Representation model (HuMoR, shown in Fig. 2), which extracts latent patterns from community-level *sequences* while making use of metadata associated with location-IDs.

HuMoR is a hierarchical Bayesian model, where $\boldsymbol{\mathcal{S}}$ is modeled as a mixture over an underlying set of motifs, and each motif is modeled as a mixture over an underlying set of location-IDs. The parameter of the motif distribution is conditioned on the sequence side features. This conditioning is intended to capture additional similarity between sequences, given features imposed by users. That is, by imposing such conditioning, motifs not only capture locations appearing together, but also locations with similar side features, even if they did not co-occur in the same sequences.

The generative process of HuMoR is:

1. For each motif, $k = 1 : K$

   (a) Draw $\boldsymbol{\phi}_k \sim \text{Dirichlet}(\beta)$

   (b) Draw $\boldsymbol{\lambda}_k \sim [\text{Normal}(\boldsymbol{0}, \sigma_0^2 I), \text{Normal}(0, \sigma^2)]$

2. For each sequence $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}$

   (a) Assign $\boldsymbol{\alpha}_{\boldsymbol{s}} = exp(\boldsymbol{x_s}^T \boldsymbol{\Lambda})$

   (b) Draw $\boldsymbol{\theta_s} \sim \text{Dirichlet}(\boldsymbol{\alpha_s})$

   (c) For each location-ID in $\boldsymbol{s}, i = 1 : \mathcal{I}$

       i. Draw $z \sim \text{Multinomial}(\boldsymbol{\theta_s})$

       ii. Draw $c_{\boldsymbol{s},i} \sim \text{Multinomial}(\boldsymbol{\phi}_z)$

where $\boldsymbol{\lambda}_k$ is a per-motif vector with length n (the cardinality of side features + 1), where $n - 1$ elements are drawn from a per-motif Gaussian distribution with variance σ_0^2 and
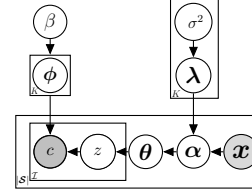

Figure 2: Graphical representation of HuMoR, where $\mathcal{I}$ is the length of a sequence, $|\boldsymbol{\mathcal{S}}|$ is the number of sequences extracted from the data, c is a location-ID drawn randomly from the location-ID distribution ϕ of motif z, $\boldsymbol{\theta}$ is the motif distribution drawn from Dirichlet($\boldsymbol{\alpha}$). $\boldsymbol{\alpha}$ is conditioned on the sequence features $\boldsymbol{x}$ and $\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is randomly drawn from a normal distribution with variance σ^2. ϕ is randomly drawn from Dirichlet(β). Nodes represent random variables. Shaded nodes are observed and plates indicate repetition.

mean 0. The nth value, a.k.a., the default value of motif k, is sampled from a Gaussian distribution with σ^2 and mean 0. To simplify notations, we refer to both variances as σ^2. $\boldsymbol{\lambda}_k, k = 1...K$ can be packed column by column, resulting in an $n \times K$ matrix denoted as $\boldsymbol{\Lambda}$. The per-motif location-ID distribution, $\boldsymbol{\phi}_k, k = 1...K$ can be packed column by column, resulting in a $|C| \times K$ matrix of location distributions, denoted as $\boldsymbol{\Phi}$. The parameter of the Dirichlet prior on the per-motif location-ID distribution is denoted by β. The parameter of the Dirichlet prior on the per-sequence motif distribution is denoted as $\boldsymbol{\alpha}$, which is a log-linear function of the sequence feature $\boldsymbol{x_s}$ and $\boldsymbol{\Lambda}$. The per-sequence motif distribution drawn from Dirichlet($\boldsymbol{\alpha_s}$), $\boldsymbol{\theta_s}, s = 1...|\boldsymbol{\mathcal{S}}|$ can be packed column by column, resulting in a $K \times |\boldsymbol{\mathcal{S}}|$ matrix denoted as $\boldsymbol{\Theta}$. z is a motif for location-ID $c_{\boldsymbol{s},i}$, i.e., the ith location-ID in sequence $\boldsymbol{s}$. K is the number of motifs in the model. In this model, c and $\boldsymbol{x}$ are the only observed variables, all other variables are latent.

The generative process of HuMoR is explained as follows. For each *sequence*, we draw a probability distribution for motifs, $\boldsymbol{\theta}$, from Dirichlet($\boldsymbol{\alpha}$), where $\boldsymbol{\alpha}$ is a function of the sequence feature $\boldsymbol{x}$ and $\boldsymbol{\Lambda}$. Then, for every element in the sequence, we randomly draw a motif z from the multinomial motif distribution, and a location-ID from the multinomial location-ID distribution of the selected motif $\boldsymbol{\phi}_z$.

Building Blocks

In this section, we discuss the two main elements that make HuMoR successful for trajectory data.
Sequences The concept of sequences is introduced to harness the power of the community in learning latent patterns for individuals. Given the variation in user activity level, which is apparent in different sources of data, it is important for models of human mobility to capture such imbalance. By introducing sequences as defined in Definition 2, the model learns from community-level sequences as opposed to individual-level trajectories. Thus, for inactive users, the risk of not learning their latent patterns is minimized as our model learns from users with similar patterns collectively.
Side Features There is more to trajectory data than

anonymized locations. HuMoR provides a general framework for adding metadata, by incorporating per-sequence side features. This approach gives an additional layer of flexibility to our model while enriching the learning process.

Example: Consider visiting time as a side feature and assume we have three time ranges, each representing a unique portion of the day (e.g., morning, evening, and night). $\boldsymbol{x_s}$ will be a vector of length 4 encoding the time feature of sequence $\boldsymbol{s}$, by containing one in the proper time range and zero otherwise in addition to a motif default value 1. As $\alpha_k = exp(\boldsymbol{x_s}^T\boldsymbol{\lambda}_k), k = 1...K$ is the parameter of per-sequence motif prior, the first $n-1$ elements in $\boldsymbol{\lambda}_k$ associate sequences having the same value of time range together, even when sequences have different location-IDs. The nth element in $\boldsymbol{\lambda}_k$ is added to result in α_k, which ultimately associates sequences having the same side features, with the same motifs. When the parameter σ^2 for drawing $\boldsymbol{\lambda}_k$ is large, the default values for all motifs will fluctuate more freely, and thus distinguish motifs from each other.

Estimating Model Parameters

In this section, we describe the algorithm used for inferring the model parameters. Note that the full joint distribution over parameters, for HuMoR in Fig. 2, has a probability distribution given by the chain rule in the following Equation,

$$P(\sigma^2, \boldsymbol{\Lambda}, \boldsymbol{x}, \boldsymbol{\alpha}, \beta, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \mathbf{z}, \boldsymbol{c}) = P(\sigma^2)P(\boldsymbol{\Lambda}|\sigma^2)\prod_{k=1}^{K}P(\boldsymbol{\phi}_k|\beta)$$

$$\prod_{s=1}^{|\boldsymbol{S}|}P(\boldsymbol{x_s})P(\boldsymbol{\alpha_s}|\boldsymbol{x_s}, \boldsymbol{\Lambda})P(\boldsymbol{\theta_s}|\boldsymbol{\alpha_s})\prod_{i=1}^{\mathcal{I}}P(z_i|\boldsymbol{\theta_s})P(c_{\boldsymbol{s},i}|\boldsymbol{\phi}_{\boldsymbol{z}}). \quad (1)$$

We are interested in reversing the generative process and learning the posterior distribution: the distribution of the latent variables $(\mathbf{z}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{\Lambda})$ given the observed data is

$$P(\boldsymbol{\Lambda}, \boldsymbol{\alpha}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \mathbf{z}|\boldsymbol{c}, \boldsymbol{x}, \beta) = \frac{P(\boldsymbol{c}, \boldsymbol{x}, \boldsymbol{\Lambda}, \boldsymbol{\alpha}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \mathbf{z}|\beta)}{P(\boldsymbol{c}, \boldsymbol{x}|\beta)}. \quad (2)$$

This distribution is intractable and cannot be computed exactly. Thus, we adopt an inference model where we alternate between estimating $\mathbf{z}$ using collapsed Gibbs sampling from the current prior distribution (i.e., sampling motif assignments) and optimizing $\boldsymbol{\Lambda}$ given $\mathbf{z}$ using the standard L-BGFS optimizer (Byrd, Nocedal, and Schnabel 1994) in MALLET (McCallum 2002). Values of $\boldsymbol{\Lambda}$ are then used in addition to the sequence side features to estimate $\boldsymbol{\alpha}$. Lastly, $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ are estimated using Equation (5) and (6), respectively. The estimation process is given in Algorithm 1.

Collapsed Gibbs sampling is an MCMC algorithm, which iteratively draws and updates one sample from the population given all other samples. In HuMoR, this implies sampling from the conditional distribution, $P(z_i|\mathbf{z}_{-i}, \mathcal{C})$, since $\mathbf{z}$ is a sufficient statistic for $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$. Sampling is given by,

$$P(z_i|\mathbf{z}^{(-i)}, \mathcal{C}) \propto (\eta_{k|s}^{(-i)} + \alpha_k) \cdot \frac{\eta_{c|k}^{(-i)} + \beta}{\eta_{.|k}^{(-i)} + \boldsymbol{\beta}}. \quad (3)$$

where $\eta_{k|s}$ is the number of times motif k was drawn in sequence $\boldsymbol{s}$, superscript $(-i)$ is used when sequence i is ex-

---

Algorithm 1: Parameter Estimation for HuMoR

> **Input:** $K, \boldsymbol{S}, \mathbf{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_{|\boldsymbol{S}|}\}, \sigma^2, \beta$
> **Output:** $\boldsymbol{\Phi}, \boldsymbol{\Theta}$
> 1   Initialize $\boldsymbol{\Lambda}, \mathbf{z}$, and update counters;
> 2   **for** *each iteration* **do**
> 3   **if** *mod(iteration,optimization interval) = 0* **then**
> 4   Optimize $\boldsymbol{\Lambda}$;
> 5   **for** *each sequence $\boldsymbol{s} \in \boldsymbol{S}$* **do**
> 6 $\boldsymbol{\alpha} = exp(\boldsymbol{x_s}^T\boldsymbol{\Lambda})$;
> 7   **for** *each location-ID $i = 1 : \mathcal{I}$* **do**
> 8   **for** *each motif $k = 1 : K$* **do**
> 9 $P(z_i|\mathbf{z}^{(-i)}, \mathcal{C}) = (\eta_{k|s}^{(-i)} + \alpha_k) \cdot \frac{\eta_{c|k}^{(-i)} + \beta}{\eta_{.|k}^{(-i)} + \boldsymbol{\beta}}$;
> 10   Sample $k \sim P(\mathbf{z}|\mathcal{C})$;
> 11 Compute $\boldsymbol{\Phi}, \boldsymbol{\Theta}$ where $\phi_{c|k} = \frac{\eta_{c|k} + \beta}{\eta_{.|k} + \boldsymbol{\beta}}$ and $\theta_{k|s} = \frac{\eta_{k|s} + \alpha_k}{\mathcal{I} + \boldsymbol{\alpha}}$;

---

cluded, $\eta_{c|k}$ is the number of times location-ID c was assigned to motif k, $\eta_{.|k} = \{\eta_{c_j|k}\}_{j=1}^{|\mathcal{C}|}$ is the total number of location-IDs assigned to motif k, and $\boldsymbol{\beta} = \sum_{c=1}^{|\mathcal{C}|} \beta$.

The location-ID distribution of each motif can be estimated as: $P(\boldsymbol{\Phi}|\boldsymbol{S}, \beta) = \frac{1}{Z_{\phi}}Dirichlet(\boldsymbol{\Phi}|\eta_{.|k} + \beta)$, where Z_{ϕ} is a normalization factor. Since the expectation of $Dirichlet(\boldsymbol{x}|\boldsymbol{\alpha})$ is

$$E(x_i) = \frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k}, \quad (4)$$

each element in the matrix $\boldsymbol{\Phi}$ is estimated using

$$\phi_{c|k} = \frac{\eta_{c|k} + \beta}{\eta_{.|k} + \boldsymbol{\beta}}. \quad (5)$$

The motif distribution is given by $P(\boldsymbol{\theta}|\mathbf{z}, \boldsymbol{\alpha}) = \frac{1}{Z_{\Theta}}Dirichlet(\boldsymbol{\theta}|\mathcal{I} + \boldsymbol{\alpha})$. Again, given Equation (4), the kth element in the motif distribution vector $\boldsymbol{\theta}$ is given by,

$$\theta_{k|s} = \frac{\eta_{k|s} + \alpha_k}{\mathcal{I} + \boldsymbol{\alpha}}. \quad (6)$$

HuMoR's complexity is $O(k \cdot |\boldsymbol{S}| \cdot l_{\boldsymbol{s}})$, which is comparable to LDA's complexity ($O(k \cdot |\mathcal{V}| \cdot l_v)$), since $|\boldsymbol{S}| \simeq (|\mathcal{V}| \cdot l_v)/l_{\boldsymbol{s}}$, where $l_{\boldsymbol{s}}$ and l_v are the average lengths of sequence and user trajectories, respectively.

Per-user Motif Distribution

As HuMoR does not model the full user trajectory generation process. We show here how to estimate the per-user motif distribution, $P(z|v)$. Let user $v \in \mathcal{V}$ has $|\boldsymbol{S}_v|$ number of sequences, after learning per-sequence motifs for all users ($\boldsymbol{\theta_s}$ in Eq. (6)), we approximate $P(z|v)$ using the chain rule,

$$P(z = k|v) = \sum_{i=1}^{|\boldsymbol{S}_v|} P(z, \boldsymbol{s}_i|v) = \sum_{i=1}^{|\boldsymbol{S}_v|} P(z|\boldsymbol{s}_i, v)P(\boldsymbol{s}_i|v). \quad (7)$$

$P(\boldsymbol{s}|v)$ can be estimated empirically from the data,

$$P(\boldsymbol{s}|v) = \prod_{i=1}^{\mathcal{I}} P(c_i|v) = \prod_{c}^{|\mathcal{C}_s|} \frac{\eta_{c|\boldsymbol{s}}}{\sum_{c'}^{|\mathcal{C}_v|} \eta_{c'|v}} = \prod_{c}^{|\mathcal{C}_s|} \frac{\eta_{c|\boldsymbol{s}}}{\eta_{.|v}} \quad (8)$$

where $\eta_{c|\boldsymbol{s}}$ and $\eta_{c|v}$ are the frequency of location-ID c in sequence $\boldsymbol{s}$ and user v, and $|\mathcal{C}_{\boldsymbol{s}}|$ and $|\mathcal{C}_v|$ are the number of unique location-IDs in sequence $\boldsymbol{s}$ and user v, respectively.

Let motif z and user v be conditionally independent given sequence $\boldsymbol{s}$, $P(z|\boldsymbol{s}, v) = P(z|\boldsymbol{s})$, using Bayes' formula and the parameters inferred by HuMoR, $P(z|\boldsymbol{s})$ is estimated by

$$P(z = k|\boldsymbol{s}) = \frac{P(\boldsymbol{s}|z)P(z)}{P(\boldsymbol{s})} = \frac{\prod_i^{\mathcal{I}} \phi_{c_i|k}\theta_{k|\boldsymbol{s}}}{\sum_{k'}(\prod_i^{\mathcal{I}} \phi_{c_i|k'})\theta_{k'|\boldsymbol{s}}}. \quad (9)$$

Therefore, taking Equations (8) and (9) into (7), we have

$$P(z = k|v) = \sum_{\boldsymbol{s}}^{|\boldsymbol{S}_v|} \left(\frac{\prod_i^{\mathcal{I}} \phi_{c_i|k}\theta_{k|\boldsymbol{s}}}{\sum_{k'}(\prod_i^{\mathcal{I}} \phi_{c_i|k'})\theta_{k'|\boldsymbol{s}}} \cdot \prod_c^{|\mathcal{C}_{\boldsymbol{s}}|} \frac{\eta_{c|v}}{\eta_{.|v}} \right), \quad (10)$$

which represents the per-user motif distribution, $\boldsymbol{\theta}_v$.

For new users, we estimate the per-user motif distribution using

$$P(z = k|v) = \sum_c^{|\mathcal{C}_v|} \frac{\phi_{c|v}}{\sum_{k'} \phi_{c|k'}} \cdot \frac{\eta_{c|v}}{\eta_{.|v}}, \quad (11)$$

which is simpler and more efficient than *folding in* approach and produces comparable results.

Evaluation

Our objective is to evaluate the proposed model and verify that the new representation preserves the latent patterns in the original raw feature space. As HuMoR is inspired by topic models, we compare our model with LDA and LDA-Asymmetric (LDA-A), which is an LDA model with asymmetric prior on the per-user motif distribution. Additionally, we compare with Principal Component Analysis (PCA) and Human Effort (HE), where features are extracted by experts based on task-dependent knowledge. It is important to note that we do not adopt conventional methods used for evaluating topic models (e.g., likelihood of held-out documents) because the likelihood of different models is not comparable (Chang et al. 2009). As a result, we evaluate our model by measuring the performance of the new representation on an application of social link prediction, where better results imply better representation, as this indicates minimum loss of information and effective latent motif expression.

We evaluate our approach on two publicly available datasets: MDC and GW. Table 2 shows summary statistics of the datasets. MDC is a CDR data, where cell-IDs associated with communication records are regarded as visited locations. Communications between users are represented as undirected links in the social graph. GW is an LBSN, where visited locations are reported by users, and friendships in the network are represented as undirected links in the social graph. Comparing to GW, MDC contains more visited locations per user. This is simply because visits are collected automatically by mobile carriers as opposed to LBSN where users manually checkin to locations.

We applied PCA, LDA and LDA-A on the same *user × location* frequency matrix. We train all graphical models for 1000 iterations, and optimize parameters every 50 iterations, after an initial burn-in period of 200. The common parameter β for all graphical models is set to be $\beta = 0.01$. For

Table 2: Datasets Statistics. No. of Nodes (No.N), No. of Edges (No.E), No. of Checkins (No.C), averaged Checkins per User (CpU) and Checkins per Location (CpL).

	No.N	No.E	No.C	CpU	CpL
MDC	107	197	1,270,045	11870	25
GW	43,956	222,319	2,887,700	65	29
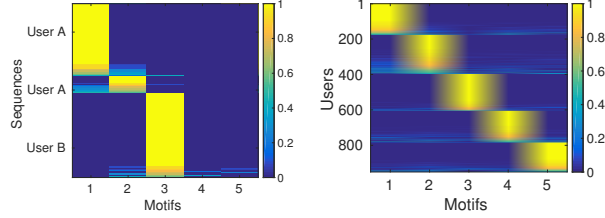
(a) $\boldsymbol{\theta}_{\boldsymbol{s}}$ for sequences of 2 users   (b) $\boldsymbol{\theta}_v$ of 950 users

Figure 3: Heatmap of per-sequence and per-user motif distribution. Note in 3(a), user A has sequences with 2 distinct patterns, while user B has only 1 pattern.

HuMoR, we use *time range* as a sequence feature, where we divided the day into four equal intervals. We have tested different intervals for time range, and found that the results are robust against the variation in the settings of this feature. The first $n - 1$ elements in $\boldsymbol{\lambda}$, associated with sequence features, are drawn from a Gaussian distribution with $\sigma^2 = 0.5$. The nth element, associated with the motif default value, is drawn from a Gaussian distribution with $\sigma^2 = 100$. Collectively, the value of $\boldsymbol{\lambda}$ along with the sequence features determine $\boldsymbol{\alpha}$, which in turn controls the per-sequence motif distribution. Sequences with similar $\boldsymbol{\alpha}$'s, will have similar motif distributions. Thus, setting a large variance for the motif default value enables motifs to be easily distinguished from each other. We study the sensitivity of σ^2 and report the results later. Regarding the number of motifs K, we study its sensitivity after comparing the performance of different models, where K is set to 5.

Fig. 3 shows a heatmap of the new representation learned by HuMoR per sequence ($\boldsymbol{\theta}_{\boldsymbol{s}}$, Fig. 3(a)) for 2 selected users and per user ($\boldsymbol{\theta}_v$, Fig. 3(b)) for 950 users. The new representation is mostly sparse, i.e., sequences (and users) are associated with only few motifs. The learned motif distribution properly captures latent patterns of sequences and users, and can instantly be used to group them into clusters.

Link Prediction

Given a subset of the network, G_{known}, which contains all nodes in G and only a subset of the edges, the task of link prediction is to infer the missing (new) links. Predicted links can either be positive, meaning a connection exists between the pair of users, or negative with no connection between users. Link prediction is a challenging problem due to the extremely unbalanced proportion of positive to negative links (3.4% positive links in MDC and only 0.02% in GW). Link prediction has been studied with unsupervised methods (Liben-Nowell and Kleinberg 2003), supervised classifica-

Table 3: Mean AUC for link prediction. Columns indicate the new representation used in conjunction with pairs of topology measures and prediction approaches (in rows). Three prediction approaches are used: Decision Trees, KNN, and Matrix Factorization (MF) on the Adjacency Matrix (AM). Column Top (for Topology) gives AUC values when using only three topology measures (CN, AA, and Jacc) in unsupervised link prediction.

Methods		MDC						GW					
		HE	PCA	LDA	LDA-A	HuMoR	Top	HE	PCA	LDA	LDA-A	HuMoR	Top
Trees	CN	0.782	0.768	0.761	0.768	**0.800**	CN	0.778	0.793	0.815	0.777	**0.912**	CN
	AA	0.779	0.759	0.769	0.768	**0.799**	0.645	0.876	0.875	0.847	0.841	**0.917**	0.892
	Jacc	0.779	0.769	0.762	0.779	**0.805**	AA	0.772	0.789	0.814	0.805	**0.892**	AA
KNN	CN	0.859	0.831	0.858	0.870	**0.891**	0.645	0.727	0.689	0.881	0.833	**0.969**	0.901
	AA	0.861	0.831	0.864	0.872	**0.888**	Jacc	0.934	0.938	0.936	0.939	**0.972**	Jacc
	Jacc	0.845	0.831	0.866	0.877	**0.893**	0.645	0.721	0.774	0.914	0.892	**0.954**	0.832
MF	AM	0.873	0.893	0.881	0.891	**0.943**	0.820	0.922	0.942	0.932	0.935	**0.950**	0.906

tion methods (Dong et al. 2012), and Matrix Factorization (MF) (Menon and Elkan 2011).

First, we use unsupervised methods based on only measures extracted from the network topology. We consider three measures: (1) the number of common neighbors between pairs of nodes v_i and v_j: $CN(v_i, v_j) = |\mathcal{N}(v_i) \cap \mathcal{N}(v_j)|$, where $|\cdot|$ gives the cardinality of a set and $\mathcal{N}(\cdot)$ is the set of 1-hop neighbors of a user; (2) Adamic-Adar (Adamic and Adar 2003), $AA(v_i, v_j) = \sum_{n \in \mathcal{N}(v_i) \cap \mathcal{N}(v_j)} \frac{1}{log(|\mathcal{N}(n)|)}$; and (3) Jaccard coefficient $Jacc(v_i, v_j) = \frac{|\mathcal{N}(v_i) \cap \mathcal{N}(v_j)|}{|\mathcal{N}(v_i) \cup \mathcal{N}(v_j)|}$. A link is predicted between user v_i and v_j when their measure is above a threshold, varying which results in an ROC curve. The Area Under the ROC Curve (AUC) using each of these measures for both datasets is listed in Table 3 under the *Topology (Top)* column.

Second, we add user representation to topology measures to learn classifiers for link prediction. For one pair of users v_i and v_j, the existence of link between them is predicted as a binary classification problem with input $\{\boldsymbol{\theta}_{v_i} - \boldsymbol{\theta}_{v_j}, Topology(v_i, v_j)\}$, where $\boldsymbol{\theta}_v$ is user v's new representation learned from trajectories and $Topology$ can be any of CN, AA or Jacc. We employ two classification approaches: Decision Trees and K-Nearest Neighbors (KNN). To overcome the issue of unbalanced labels, we limit link prediction to the 2-hop neighbors of each user. Thus, eliminating many true negative links from the learning space. Table 3 shows the AUC averaged over 10 runs when adding five different representation produced by HE, PCA, LDA, LDA-A and HuMoR to the three different topology measures. In both datasets when comparing to the column of *Topology*, we see that the new representation learned by HuMoR always improves the results significantly and outperforms all other representation learning methods. Some methods even worsen the results because of the inappropriately added representations.

Finally, we do link prediction using MF with and without the learned representations. Table 3 shows that adding representations always improves the results, and the best results are achieved by HuMoR. The overall best result in MDC is achieved by MF with representations from HuMoR. In GW, the best result is achieved by KNN with representations from HuMoR added to AA. All the results indicate that the new representation learned by HuMoR is better than the others.

When comparing the results achieved using HuMoR on both datasets, we note that profound improvement is obtained in the GW dataset. Recall that GW is an LBSN dataset with a large number of inactive users, who mostly have sparse location traces. As HuMoR captures patterns from collective sequences of all individuals in the dataset, the sparsity of location traces does not impact the model's performance as it does with mainstream topic models (LDA and LDA-A).

Parameter Sensitivity

We study the sensitivity of HuMoR to the setting of the number of motifs (K) and to the motif variance (σ^2) in Fig. 4. Fig. 4(a) shows the AUC results for both datasets when varying the value of K. For the MDC data, prediction results degrade when the number of motifs increases over 5. This is not the case with the GW data as no significant change is observed. The main reason is that MDC has a small number of users for which a small K is sufficient for capturing the latent pattens. Fig. 4(b) shows that the model is robust to the setting of σ^2.
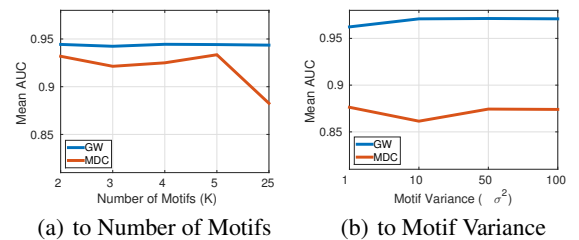

(a) to Number of Motifs   (b) to Motif Variance

Figure 4: Sensitivity analysis of HuMoR parameters.

Conclusions and Future Work

We propose HuMoR, a PGM designed to infer latent patterns from anonymized location traces with power-law distributed locations. Unlike mainstream topic models, HuMoR is capable of learning patterns for users with different activity levels and including side features in the learning process. We admit that HuMoR's efficiency is limited by Gibbs sampling. Thus adopting an efficient parameter inference technique is a main direction for future work.

Acknowledgments

References

Adamic, L. A., and Adar, E. 2003. Friends and neighbors on the web. *Social Networks* 25(3):211–230.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR* 3:993–1022.

Byrd, R. H.; Nocedal, J.; and Schnabel, R. B. 1994. Representations of quasi-newton matrices and their use in limited memory methods. *Math. Prog.* 63(1-3):129–156.

Chang, J.; Gerrish, S.; Wang, C.; Boyd-Graber, J. L.; and Blei, D. M. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*.

Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and mobility: user movement in location-based social networks. In *KDD*.

Dong, Y.; Tang, J.; Wu, S.; Tian, J.; Chawla, N. V.; Rao, J.; and Cao, H. 2012. Link prediction and recommendation across heterogeneous social networks. In *ICDM*.

Eagle, N., and Pentland, A. S. 2009. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology* 63(7):1057–1066.

Farrahi, K., and Gatica-Perez, D. 2011. Discovering routines from large-scale human locations using probabilistic topic models. *TIST* 2(1):3.

Gonzalez, M. C.; Hidalgo, C. A.; and Barabasi, A.-L. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–782.

Gruber, A.; Weiss, Y.; and Rosen-Zvi, M. 2007. Hidden topic markov models. In *AISTATS*.

Idé, T., and Sugiyama, M. 2011. Trajectory regression on road networks. In *AAAI*.

Joseph, K.; Tan, C. H.; and Carley, K. M. 2012. Beyond local, categories and friends: clustering foursquare users with latent topics. In *UbiComp*.

Kapicioglu, B.; Rosenberg, D. S.; Schapire, R. E.; and Jebara, T. 2015. Collaborative place models. In *IJCAI*.

Laurila, J. K.; Gatica-Perez, D.; Aad, I.; Bornet, O.; Do, T.-M.-T.; Dousse, O.; Eberle, J.; Miettinen, M.; et al. 2012. The mobile data challenge: Big data for mobile computing research. In *PerCom*.

Liben-Nowell, D., and Kleinberg, J. 2003. The link-prediction problem for social networks. In *CIKM*.

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Menon, A. K., and Elkan, C. 2011. Link prediction via matrix factorization. In *ECML/PKDD*.

Mimno, D. M., and McCallum, A. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*.

Phung, D.; Adams, B.; and Venkatesh, S. 2008. Computable social patterns from sparse sensor data. In *LocWeb*.

Schumann, A., and Lécué, F. 2015. Minimizing user involvement for accurate ontology matching problems. In *AAAI*.

Song, C.; Qu, Z.; Blumm, N.; and Barabási, A.-L. 2010. Limits of predictability in human mobility. *Science* 327(5968):1018–1021.

Tang, J.; Meng, Z.; Nguyen, X.; Mei, Q.; and Zhang, M. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *ICML*.

Toole, J. L.; Ulm, M.; González, M. C.; and Bauer, D. 2012. Inferring land use from mobile phone activity. In *KDD*.

Wallach, H. M.; Mimno, D. M.; and McCallum, A. 2009. Rethinking LDA: Why priors matter. In *NIPS*.

Wallach, H. M. 2006. Topic modeling: beyond bag-of-words. In *ICML*.

Wang, X.; McCallum, A.; and Wei, X. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM*.

Yuan, J.; Zheng, Y.; and Xie, X. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *KDD*.

Zheng, Y.; Zhang, H.; and Yu, Y. 2015. Detecting collective anomalies from multiple spatio-temporal datasets across different domains. In *SIGSPATIAL*.

Zheng, Y. 2015a. Methodologies for cross-domain data fusion: An overview. *IEEE Transactions on Big Data*.

Zheng, Y. 2015b. Trajectory data mining: An overview. *TIST* 6(3):29.