An Unsupervised Framework for Sensing Individual and Cluster Behavior Patterns From Human Mobile Data

Jiangchuan Zheng[♮] **and Lionel M. Ni**[♮,♯]
[♮]Department of Computer Science and Engineering
[♯]Guangzhou HKUST Fok Ying Tung Graduate School
Hong Kong University of Science and Technology
{jczheng,ni}@cse.ust.hk

ABSTRACT
Human behavior understanding is a fundamental problem in many ubiquitous applications. It aims to automatically uncover and quantify characteristic behavior patterns in users' daily lives as well as disclose behavior clustering structure among multiple users. The key challenge is how to define a naturally interpreted representation for users' daily behavior patterns, which can be easily exploited to not only uncover the behavior similarity among multiple users but also predict users' future activities. In this paper, we define such a representation, and propose a probabilistic framework which can automatically learn it from mass amount of mobile data in unsupervised setting and exploit it to predict user activities. By an appropriate information sharing among multiple users, this framework overcomes single-user data sparsity problem and effectively identifies behavior clustering structures in a set of users. Experiments conducted on a public reality mining data set demonstrate the effectiveness and accuracy of our methods.

Author Keywords
Human Behavior Learning, Mobile Phone Sensing, Human Activity Inference, Graphical Models

ACM Classification Keywords
I.5 Pattern Recognition; H.5.2 User/Machine Systems

General Terms
Algorithms, Experimentation

INTRODUCTION
Recent years have witnessed a wide applicability of mobile phones in humans' daily lives. In addition to providing abundant social network data through calls made among people which is valuable for the study and inference of social interaction patterns [17, 18], mobile phones can also be treated as wearable sensors, which generate sequences of spatial-temporal signals that help us understand the behavior patterns of individuals such as their habits and characteristic activities as well as the structure and dynamics of social system [5, 16]. This understanding serves as a basis for many upper-level intelligent applications such as social relationship inferences [10] and location-based services. From ubiquitous computing's perspective, given mass amount of timestamped location data generated by many people's mobile devices, we are interested in the following questions:

1. What is the internal behavioral pattern of each person, and how to discover and quantify that pattern automatically?

2. How to uncover the internal behavior similarity and differences among multiple persons?

3. How to predict people's future activities using the behavior patterns learned from historical data and to what extent can they be predicted?

In this paper, we aim to develop a unified probabilistic framework that can address these questions collectively. We firstly describe what a behavior pattern is in our context as it is the key concept in these questions. For humans, the behavior pattern can be thought of as regular temporal transitions between typical states. For example, a person who leads an orderly life might usually go to work at about 8am and come home before 6pm; a graduate student, who has more freedom, might always sleep very late in the morning until 11am, and then work till midnight. Home and work are two typical states involved in these examples and in most other cases, but the inclusion of other typical states such as doing physical exercises regularly in one's life is also possible. Typically, human's states are associated with relatively fixed locations [14], which can be obtained from the mobile data collected, though noises might exist. Empirical studies made in [7, 8] show that most humans lead low-entropy lives, that is, their lives exhibit strong regularity in the long term. From this perspective, we can say that a person's characteristic behavior pattern is determined by his inherent habits and social identity. This fact makes it meaningful to build statistical models to learn internal behavior patterns hidden in human-generated mobile sensor data and make use of that knowledge to predict human's future behaviors.

This paper shows that by establishing correlation between spatial and temporal dimensions, certain aspects of humans' routine behaviors can be revealed, which provide us with

UbiComp '12, Sep 5-Sep 8, 2012, Pittsburgh, USA.

useful insights into both individual and group. At individual level, knowing a person's behavior pattern might help us infer his/her profession or work type. At group level, knowing the similarity and differences in behavior patterns among multiple persons might help us infer their friendship or social organization membership. For example, students who are in the same department or students who are close friends tend to behave similarly, though an accurate inference of such relationship also depends on other social information. Therefore, when studying a set of users, it is tempting to cluster them according to their behavior patterns, that is, identify a few most characteristic latent behavior patterns which can help us divide these users into several groups such that users in the same group behave similarly and users from different groups have noticeable behavior differences, rather than studying each user separately. Such a group division usually implies certain latent social structures, which is valuable information for many applications.

However, automatic discovery of meaningful human behavior patterns from mass amount of mobile data is not a trivial task for the following reasons. First, noises and sparsity are present in the raw data collected. For noises, the location information reported, which are usually in the form of cell tower id, is often inaccurate due to the sparsity of cell towers, and can change frequently among several neighboring cell towers even when the user stays at a fixed position. Also, cell tower id's are not associated with the semantics of human's latent states such as "Home" or "Work". For sparsity, the sequence of mobile sensor records is usually not time-continuous, as phones might be powered off from time to time. Also, the data collected for some specific users might be scarce, which require us to exploit the data of other users who behave similarly(typically friends or in the same organization) to assist learning those users' behaviors, which is technically challenging. Second, choosing a reasonable statistical representation for humans' behavior pattern is not an easy task. This representation should be easy for learning, should support effective and efficient measurement of behavior similarity and differences among people thus achieving behavior clustering, and should be easy to exploit to make probabilistic inference for people's future activities. Third, identifying a few most characteristic latent behavior patterns to achieve behavior clustering requires us to exploit all users' behavior data in a principled collaborative way rather than studying each user separately, which poses technical challenges. To tackle these challenges, we make the following contributions in this paper:

1. We define a probabilistic representation for humans' behavior patterns which not only captures the most informative features but also well encodes the uncertainty in their behaviors. This representation can also be well exploited to discover behavior cluster structure as well as predict human's future activities in a probabilistic way.

2. We propose and develop a probabilistic framework, specifically, a Bayesian network, to learn single user's latent behavior patterns from mass amount of raw data, which can effectively tackle the problems of sparsity and noises. This framework is unsupervised in nature as it does not require that the state semantics of cell towers be pre-labeled.

3. We propose algorithms to uncover humans' behavior cluster structures. Specifically, we extend the previous Bayesian network which only models a single user to a hierarchical one which explicitly models the behavior similarity and differences among multiple users, and for each user learn the likelihood that he adopts each typical latent behavior pattern. This method overcomes the single-user data scarcity problem by exploiting the information of other users to assist the learning of a single user's behavior pattern.

4. We apply standard inference techniques to predict humans' future locations based on time using the behavior patterns learned. This prediction is probabilistic in nature, which can also answer the question of to what extent a person's location can be predicted at each time point by exploiting the uncertainty encoded in the learned behavior patterns.

DATA

The data set we used to validate our model is a public Reality Mining data set [1] collected by MIT Media lab on 95 academic mobile phone users over approximately 9 months. In contrast to other traditional mobile phone data sets, this one has the following special characteristics, which renders it extremely suitable for our study.

1. Special software has been pre-installed on the phones used by the studied users such that each time a user changed cell tower, the id of this new serving tower as well as the current time will be immediately recorded as long as the phone is powered on, which means that we can continuously follow the users throughout the day. This property significantly diminishes the severe sparsity problem encountered in traditional mobile phone data sets, where location information is recorded only when the user makes a phone call, in which case learning users' behavior patterns is somewhat technically infeasible. But sparsity problem still exists as phones might be powered off or users might forget to bring their phones with them from time to time.

2. As the studied users are all academic users from MIT including both students and staff, their daily behaviors are highly regular compared with traditional mobile phone data generated by operators. This makes the study and learning of users' behavior patterns more meaningful and verifiable.

3. There exists clear group or organization division in this data set; some users come from MIT Media Lab, and others come from Sloan Business School, including both students and staff members. This underlying structure provides some ground truth for validating the behavior clustering results generated by our model, as users' personal behavior patterns are usually reflective of organizational rhythms [2].

4. This data set is accompanied by some survey data, which reveals much useful information such as what are the typical working hours of each user, which organization or research group a user belongs to, to what extent a user's behavior is

predictable, and who are friends to each other, etc. Such information is disclosed by the users studied, and hence can be used as ground truth to validate some of the results generated by our unsupervised learning model.

Some work [20, 16] have modeled humans' transportation modes, which is a fine-grained aspect of humans' behaviors, by exploiting sequences of time-stamped locations with spatial coordinates. We show that by using only coarse-grained time-stamped cell tower sequences without spatial coordinates involved, as is the case for this data set, certain aspects of humans routine behavior patterns can still be modeled.

RELATED WORK

The idea of uncovering regular rules in human's behaviors is not new, and some prior work [6, 9, 11] has explored this territory on the same reality mining data set. In this section, we will briefly review major previous work on this data set and pinpoint our differences compared with them.

Research in [9] had for the first time explored this data set to disclose underlying structures in humans' behaviors. Its basic idea was to apply Principle Component Analysis(PCA) to a user's mobile data over many days, with the data in each day represented as a 24-dimensional vector indicating the state of that user in each hour, to disclose his characteristic behavior called eigenbehavior such as commuting to work at 8:00am and staying at work until 17:00pm when returning home. The eigenbehaviors are encoded in the top eigenvectors found by PCA, with the weight values indicating the importance and correlation of the original features. Though such eigenbehavior representation can disclose users' behavior structures to some extent, this method has several drawbacks. First, PCA is a general-purpose dimensionality reduction method, which does not exploit domain-specific knowledge, that is, the special structures in humans' behaviors characterized as relatively stable transitions among typical states. Discovering humans' behavior patterns simply by assigning physical meanings to the weight values in eigenvectors is difficult and might be too subjective, as eigenvectors themselves have no well-defined physical meanings and are only used to project data to low-dimensional space. In view of this, it would be better to explicitly model the special internal structures in human behavior data, which will make the learning results more interpretable. Second, the eigenbehavior representation does not encode the uncertainty in humans' behaviors, that is, it does not tell us how certain the user is at one specific state at every time point, and hence is not faithful to reality. Third, this method studies each user separately, and has not exploited all users' information to discover the most characteristic latent behavior patterns, nor has it performed clustering to uncover behavior similarity and differences. Eigenvectors themselves are not suitable for clustering as this operation lacks theoretical support. Fourth, this method cannot be used to predict humans' future locations, not to mention probabilistic prediction. Our proposed model aims to compensate these drawbacks.

Another work [11] applies the latent topic model LDA(Latent Dirichlet Allocation) [4] to discover latent characteristic routine topics of humans from the same data set by exploiting all users' data. Typical routine topics they have found are "going to work at 10am","staying at home for the entire evening". Although such common topics are meaningful and characteristic, they cannot tell us the characteristic behavior pattern of a specific user at daily granularity such as "go to work at about 8am and come home before 6pm almost every day", nor can it disclose the daily behavior similarity and differences among multiple users. Also, its approach of encoding time and state-labeled location together as "activity words" couples spatial and temporal features. As a result, it cannot achieve time-based user location prediction, nor can it exploit the continuous semantic of time, suffering from sparsity problem. Furthermore, this model is supervised, because it requires that cell tower id's be labeled in advance with state semantics such as "Home", "Work" for each user before learning can be performed, which is not practical in real world as such information can hardly be obtained, especially in large-scale study. Our model attempts to overcome these limitations.

The third work [6] applies the concept of conditional entropy in information theory to the same data set, assessing for each typical time point, to what extent can a user's future activity be predicted by the information at that time point. However, it does not address how to make that prediction when the predictability is high. In contrast, our model can make formal predictions and can also address the predictability issue using the knowledge learned which encodes uncertainty.

MODEL

This section elaborates the design of our probabilistic framework and illustrates how it tackles the three questions we posed at the very beginning collectively and overcomes the challenges we mentioned. We firstly define some basic notations. Suppose there are N users in the mobile data set, and each user's phone reports sequences of time-stamped location records for M days. For user n, the sequence of records he reports at the mth day is denoted as $\{(t_{nmr}, x_{nmr})\}_{r=1..R}$, where t_{nmr} is a time point and x_{nmr} denotes the unique id of the cell tower where the user is located at that time point, and R is the maximum total number of records in a day.

One key step in our work is to define a reasonable probabilistic representation, also called a fingerprint, for a user's individual behavior pattern. Informally, this representation should have the following characteristics: (1) capture the most informative and important features in a user's individual behavior pattern while at the same time encoding the uncertainty of his behavior; (2) can be easily and naturally interpreted; (3) can be learned effectively and efficiently from mass amount of data; (4) can be exploited to effectively uncover the internal behavior similarity and differences among multiple users under well-defined behavior similarity metric; (5) can be exploited to make probabilistic inference for users' future activities with uncertainty involved. Defining such a representation is not trivial. To the best of our knowledge, no prior work has ever given a formal representation for human's behavior pattern which achieves these characteristics. Defining an informative representation for users'

behavior pattern and learning that pattern from raw mobile data is essentially a dimensionality reduction process, which maps the mass amount of raw spatial-temporal features reported by the user to a latent space in which his characteristic behavior is revealed. Compared with other general-purpose dimensionality reduction methods such as PCA which is used by [9] to uncover eigenbehaviors, our representation has fully exploited the special structures in humans' regular behaviors, which can be regarded as important domain-specific knowledge. From this perspective, our learning process essentially performs human behavior-oriented dimensionality reduction, which we believe will be more faithful to this specific scenario than other general-purpose methods.

Single-user Model

Since the daily behavior of a single user leading a low-entropy life can be characterized as a relatively stable temporal transition among his typical states such as "home","work", we define the daily behavior pattern of a user as a set of temporal distributions associated with each typical state. Intuitively, for a user, his different states are usually characterized by different temporal distributions. In literature, Hidden Markov Model is usually used to model and learn state transition probabilities. However, state transition is not what we are concerned about in this task. We are more concerned about the state-conditioned temporal characteristics, as it is the set of temporal distributions associated with each typical state that characterizes a user's daily behavior pattern. Thus, we define a user's daily behavior pattern as follows:

Definition 1: A user's daily behavior pattern is defined as a set of temporal distributions associated with each of his typical states: $\{P(t|s)|s \in S\}$, where S is a set containing all possible states, and $P(t|s)$ is the associated temporal distribution when this user is at state s in a day.

The physical meanings of the states in S are application-specific. Some typical ones are "at home", "at work", "having dinner", but the inclusion of other typical states for a specific user does not change this definition as well as the learning process which will be illustrated later. The form of the conditional probability density function $P(t|s)$ is temporarily not explicitly stated, thus, this definition is sufficient to describe the behavior pattern of a user who lives low-entropy lives, in which case each of his typical states has a relatively fixed temporal distribution in the long term. For example, a user who usually works from 11am to 8pm and spends most other time at home can be characterized by two temporal distributions, one is associated with "work" state which assigns higher probability to the time points in the afternoon, the other is associated with "home" state which assigns higher probability to the time points in the morning and evening. However, to enable automatic learning of users' latent behavior patterns, we need to explicitly define $P(t|s)$. As t is continuous in this scenario, we naturally choose one-dimensional Gaussian distribution, as its exponential nature will make the subsequent learning process tractable and efficient. Yet, a single Gaussian is not sufficient to describe the temporal distribution of a state, as one typical state might be scattered in two or more non-continuous time intervals. For instance, a user might work in both morning and afternoon, but stays at home for 2 or 3 hours at noon. Taking this fact into account, for the sake of generality, we model state-conditioned temporal distribution $P(t|s)$ as a Gaussian mixture, namely $P(t|s) = \sum_{T=1}^{H} P(t|T,s)P(T|s)$. In this model, there can be up to H temporal Gaussian components associated with one state s, each of which is indicated by the latent variable T. $P(T|s)$ indicates the proportion of each temporal Gaussian component with respect to state s.

This behavior representation is probabilistic in nature, and boasts the 5 characteristics mentioned above. First, it captures the most informative features in a user's behavior pattern through the mean and variance parameters in each Gaussian component of each state, which together describe a temporal interval in a probabilistic way encoding uncertainty. Second, for a specific state s, $P(t|s)$ can be interpreted naturally and easily as behavior pattern, with its Gaussian components specifying approximately when the user is at state s in his daily life. This contrasts sharply with the behavior representation in [9]. Third, the Gaussian parameters in $P(t|s)$ for each s can be effectively and efficiently learned when combined with cell tower information and incorporated into a graphical model, which will be elaborated later. Fourth, since users' behavior patterns are finally characterized as a set of distributions, behavior similarity and differences among multiple users can be easily evaluated as there exists well-studied metric in machine learning literature measuring the "distance" between different distributions, which will make the task of behavior clustering feasible and theoretically sound. Fifth, probabilistic inference for users' future activities can be made under this model by applying standard inference algorithms in graphical model theory, which will be illustrated later.

Until now, we have only defined a probabilistic representation for users' regular behavior patterns. However, since state s is not known in the raw mobile data, i.e., each record in the raw mobile data is not labeled with user's state information such as "home", "work", direct learning makes no sense. Despite of this, cell tower id's are recorded with time points in the data. Although the recorded cell tower id's might be inaccurate or changing frequently due to the sparsity of cell tower locations, one observation is that when a user stays at a fixed location for a relatively long period, which usually implies a latent state, the nearby cell towers will be recorded with much higher probability than other farther cell towers. Also, different states of a user are usually associated with apparently different locations. This prompts us to make use of those uncertain cell tower information to infer which mobile records in a user's data are likely to belong to one latent state so that the time points associated with different latent states can be approximately separated to uncover the latent behavior pattern of a user. By combining these ideas together, we give the template Bayesian network representation in figure 1, which is the model we use to uncover a single user's latent behavior pattern.

Suppose M days' mobile data are collected for a specific user and each day contains up to R records. Each of those

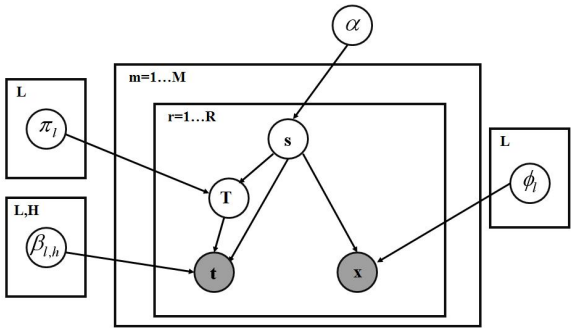

Figure 1. Graphical representation of single-user behavior model

records contains two attributes, time and cell tower id, denoted as t and x, respectively, which are observed variables. As discussed before, humans' behavior patterns are characterized as regular transitions between typical states such as "home","work","having dinner",etc, so each record (t, x) is associated with a hidden variable s denoting the latent state of this user at that time point t, whose value is unknown to us, hence making the model purely unsupervised. Assume that there are L possible states. The behavior representation $\{P(t|s)|s = 1, 2, ..., L\}$ we want to learn, namely the set of state-conditioned temporal distributions, is encoded as a set of Gaussian mixtures where each Gaussian mixture corresponds to one state $s = l$, with T as the latent Gaussian component indicator variable. π_l, $\{\beta_{l,h}\}_{h=1...H}$ are parameters governing H Gaussian components associated with state $s = l$, where $\pi_l = \{\pi_{lh}\}_{h=1..H}$ with $\sum_{h=1}^{H} \pi_{lh} = 1$ indicates the proportion of each Gaussian component w.r.t state $s = l$, and $\beta_{l,h} = \{\mu_{l,h}, \lambda_{l,h}\}$ specifies the mean and precision(inverse variance) of the hth Gaussian component associated with state $s = l$, that is, $P(t|s = l, T = h) \sim \mathcal{N}(\mu_{l,h}, \lambda_{l,h}^{-1})$. α encodes the probability at which this user adopts each state, for example, α may reveal such knowledge that a user usually spends 80% time at work and 20% time at home. Essentially, α,$\{\pi_l\}_L$, $\{\beta_{l,h}\}_{L,H}$ together characterize the temporal distribution for each typical state of the user, which will be learned from data. In addition to being characterized by temporal distribution, a hidden state is also characterized by spatial distribution, that is, the cell tower distribution, which assigns high probability to those cell towers near the user's typical location when he is at that state. In fact, it is those cell tower information that helps the learning algorithm infer the possible hidden state for each record, thus separating time points associated with different states so that the underlying behavior pattern can be uncovered. The basic observation utilized in this process is that each typical state is associated with a relatively fixed cell tower distribution, which can be seen as a spatial characterization for that state. The state-conditioned cell tower distribution is encoded as the conditional probability distribution $P(x|s)$ in this graphical model, with the parameter ϕ_l encoding the knowledge of which cell towers are most likely to be associated with state $s = l$, that is, it governs the conditional distribution $P(x|s = l)$. Since in this reality mining data set, the cell tower information provided is only in the form of discrete id's rather than continuous spatial coordinates, we cannot exploit the spatial closeness relationship between neighboring cell towers by modeling $P(x|s = l)$ as a two-dimensional Gaussian. To deal with this situation and adapt to the discrete nature of cell tower id, we model $P(x|s = l)$ as a conditional multinomial distribution $P(x|s = l) \sim Mul(x|\phi_l)$. Suppose there are F cell towers in total, then $\phi_l = \{\phi_{lf}\}_{f=1..F}$ with $\sum_{f=1}^{F} \phi_{lf} = 1$, and ϕ_{lf} indicates the probability that the cell tower with id f is recorded when the user is at his lth state. Essentially, $\{\phi_l\}_L$ characterize the spatial distribution for each typical state of the user, which will be learned from data.

In this model, the hidden state s serves as a bridge connecting temporal and spatial features of each record together. For a user, a hidden state is characterized by both temporal and spatial distributions, for example, when the user is at home, the current time usually falls in certain fixed time periods, say, early morning or evening, and the cell towers recorded are most likely the few ones around his home although arbitrary change among them might happen. Under this model, the learning process will systematically make use of both temporal and spatial information to infer the hidden state of each record in a principled way, not only uncovering the behavior pattern of the user, but also revealing the user's major locations at each state. In fact, the time and cell tower information "reinforce" each other during the learning process, that is, it alternates between using cell tower id's to infer the state semantics of time points and using time points to infer the state semantics of cell tower id's until convergence. One advantage of this method over previous work is that it is purely unsupervised, without requiring users to label state semantics for major cell towers as done in [9] and [11].

This model can only learn the behavior of a single user, and is a specific case of the multi-user model. So the illustration of the learning process will be delayed to the next section.

Multi-user Model

In this section, we extend the single-user model to a multi-user model, which explicitly models the behavior similarity and differences among multiple users, aiming at uncovering the underlying behavior cluster structure and answering question 2 raised at the beginning of this paper. Intuitively, although the behavior details vary from user to user, there always exists a few typical behavior patterns by using which each user's behavior can be characterized appropriately. As an example, in this data set, students from MIT Media Lab and students from Sloan Business School are very likely to adopt apparently different daily behavior patterns, for instance with different working hours, which are due to rhythm differences in different organizations. Therefore, when studying a set of users, it is more useful to identify a few typical behavior patterns among these users, which can help us infer the profession, identity division of these users as well as social structures, than studying each user separately, which might generate many redundant non-informative patterns and suffer from over-fitting problem.

To extend the single-user model to a multi-user model, we combine multiple users' data together, separate different days,

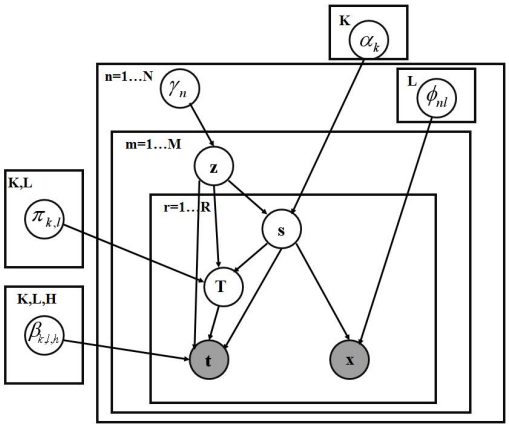

Figure 2. Graphical representation of multi-user behavior model

and introduce a hidden variable z for each day of each user, resulting in the hierarchical Bayesian network in figure 2. In this hierarchical Bayesian framework, a behavior pattern is still represented as a set of state-conditioned temporal distributions with each state inferred from cell tower distributions as before, so the structure of the most inner layer remains unchanged. The difference, compared with the single-user model, is that rather than learning for each user a specific behavior pattern, we learn K (K is small) typical behavior patterns hidden in the target set of users(similar to the idea of "centroids" in traditional clustering problem) and represent each user's behavior as a probabilistic mixture of those K patterns. This process essentially discovers the latent behavior clustering structure by exploiting the information of all users in a principled way, thus overcoming the sparsity problem in certain users' data. In this model, there are K possible values for z, representing K typical behavior patterns. The kth typical behavior pattern is encoded in parameters α_k,$\{\pi_{k,l}\}_L$,$\{\beta_{k,l,h}\}_{L,H}$, whose meanings are similar to what is stated in the section of single-user model, but are with respect to a "centroid" behavior rather than a specific user's behavior. We put the hidden variable z in the "day" layer to account for the variation of a user's behavior over days. Each user is associated with a vector parameter $\gamma = \{\gamma_k\}_{k=1..K}$ encoding at what probability this user adopts each typical behavior pattern in his life, which can be used to compare behavior similarity among multiple users. Essentially, each user is represented as a probabilistic mixture of those K typical behavior patterns which are found by exploiting all users' information, thus overcoming the single-user sparsity problem. The vector parameter ϕ resides in the "user" layer as different users usually have different locations even for the same state. In some sense, the multi-user model can be viewed as a variant of Author-Topic model(ATM)[11], with the difference that in our specific scenario, a latent topic refers to a typical behavior pattern, and is characterized by a set of state-involved spatial-temporal distributions rather than a simple discrete distribution such as word distribution in ATM.

We end this section by summarizing the generative process for the nth user's data under this multi-user behavior model:

- Draw γ_n from a common Dirichlet prior $\gamma_n \sim Dir(\xi)$.
- For each day m of M days:
  - Draw behavior pattern $z \sim Mul(\gamma_n)$
  - For each record r of R records:
    - Draw latent state $s \sim Mul(\alpha_z)$
    - Draw Gaussian component $T \sim Mul(\pi_{zs})$
    - Draw time point $t \sim \mathcal{N}(t|\mu_{zsT}, \lambda_{zsT}^{-1})$
    - Draw cell tower id $x \sim Mul(\phi_{ns})$

LEARNING ALGORITHM

As using only one user's data and setting $K = 1$ degenerates the multi-user model to single-user model, we only present learning algorithms for the multi-user model for generality.

EM(Expectation-maximization) is a widely adopted MLE method to learn graphical model with latent variables [3, 15]. However, since there are many latent variables in our model, there will be potentially many local maximums and EM is very likely to converge to certain uninformative local maximum. To avoid such cases, we instead adopt Bayesian learning method [15] to learn the posterior expectation of each parameter given data. To do that, we firstly introduce conjugate priors with known hyperparameters for each parameter. For multinomial parameters, the conjugate prior is Dirichlet [3]. Specifically, $\gamma_n \sim Dir(\xi)$, $\alpha_k \sim Dir(\delta)$, $\pi_l \sim Dir(\varepsilon)$, $\phi_{nl} \sim Dir(\eta)$. For Gaussian, the conjugate prior is the normal-gamma [3], $(\mu, \lambda) \sim \mathcal{N}(\mu|\mu_0, \frac{1}{\nu\lambda})Gam(\lambda|a, b)$. We use collapsed Gibbs sampling method [15] to perform Bayesian learning for the sake of its efficiency. For notation simplicity, denote all observed variables as D, all hidden variables as Z, and all parameters as θ. Collapsed Gibbs sampling learns the posterior expectation of θ given the observed data through the following formula [15]:

$$E_{\theta\sim P(\theta|D)}[\theta] = \sum_Z P(Z|D)E_{\theta\sim P(\theta|Z,D)}[\theta] \qquad (1)$$

using stochastic sampling method. Firstly, it obtains a sampling engine which is used to sample Z from $P(Z|D)$ by performing Markov Chain Monte Carlo method(MCMC) [15] until convergence. Specifically, it continuously samples each Z_i given the sampled values of all other hidden variables using the conditional probability distribution $P(Z_i|D, Z_{\setminus i})$ where the parameters have been integrated until convergence. Then it estimates $E_{\theta\sim P(\theta|D)}[\theta]$ by evaluating $E_{\theta\sim P(\theta|Z,D)}[\theta]$ for each Z sampled from $P(Z|D)$ and then compute the average. Once Z is given, $E_{\theta\sim P(\theta|Z,D)}[\theta]$ can be evaluated in closed form due to the usage of conjugate priors in our models [3], the details of which will be omitted to save space.

We now give the expression for each conditional probability distribution $P(Z_i|D, Z_{\setminus i})$ we have derived under the setting of our model.

1. Sample T_{abc}. Assume now $z_{ab} = k$, $s_{abc} = l$.

$$P(T_{abc} = h|\mathbf{t}, \mathbf{x}, \mathbf{T}_{\setminus abc}, \mathbf{s}, \mathbf{z}) \propto \qquad (2)$$
$$\frac{N_{z_{nm}=k,s_{nmr}=l,T_{nmr}=h,\setminus abc} + \varepsilon}{\sum_h N_{z_{nm}=k,s_{nmr}=l,T_{nmr}=h,\setminus abc} + H\varepsilon} \cdot$$
$$P(t_{abc}|\{t_{nmr}|z_{nm} = k, s_{nmr} = l, T_{nmr} = h, \setminus abc\})$$

2. Sample z_{ab}. Assume now $s_{abc} = l$,$T_{abc} = h$.

$$P(z_{ab} = k|\mathbf{t}, \mathbf{x}, \mathbf{T}, \mathbf{s}, \mathbf{z}_{\backslash ab}) \propto \frac{N_{z_{am}=k,\backslash b} + \xi}{\sum_k N_{z_{am}=k,\backslash b} + K\xi} \cdot P(\{s_{abr}\}|\{s_{nmr}|z_{nm} = k, \backslash ab\}) \cdot \prod_{l=1}^{L} P(S_1|S_2) \prod_{l=1}^{L} \prod_{h=1}^{H} P(S_3|S_4) \quad (3)$$

where $S_1 = \{T_{abr}|s_{abr} = l\}$,$S_2 = \{T_{nmr}|z_{nm} = k, s_{nmr} = l, \backslash ab\}$, $S_3 = \{t_{abr}|s_{abr} = l, T_{abr} = h\}$,$S_4 = \{t_{nmr}|z_{nm} = k, s_{nmr} = l, T_{nmr} = h, \backslash ab\}$.

3. Sample s_{abc}. Assume now $z_{ab} = k$,$T_{abc} = h$, $x_{abc} = f$.

$$P(s_{abc} = l|\mathbf{t}, \mathbf{x}, \mathbf{T}, \mathbf{s}_{\backslash abc}, \mathbf{z}) \propto \frac{N_{z_{nm}=k,s_{nmr}=l,\backslash abc} + \delta}{\sum_l N_{z_{nm}=k,s_{nmr}=l,\backslash abc} + L\delta} \cdot \frac{N_{z_{nm}=k,s_{nmr}=l,T_{nmr}=h,\backslash abc} + \varepsilon}{\sum_h N_{z_{nm}=k,s_{nmr}=l,T_{nmr}=h,\backslash abc} + H\varepsilon} \cdot \frac{N_{s_{amr}=l,x_{amr}=f,\backslash abc} + \eta}{\sum_f N_{s_{amr}=l,x_{amr}=f,\backslash abc} + F\eta} \cdot P(t_{abc}|\{t_{nmr}|z_{nm} = k, s_{nmr} = l, T_{nmr} = h, \backslash abc\}) \quad (4)$$

The derivation details will be omitted here to save spaces. The index abc refers to the cth record in the bth day of the ath user, while r means iterating all possible r from 1 to R and similar meanings apply to n and m. $\backslash abc$ means excluding the record indexed by abc. $N_{z_{nm}=k,s_{nmr}=h,\backslash abc}$ is the total number of records indexed by all possible n,m,r satisfying $z = k$,$s = l$ and $T = h$ excluding the one indexed by abc. Other such terms are similarly explained. The term $P(t_{abc}|S)$ where S is a set containing time points satisfying certain conditions is the posterior predictive distribution of univariate Gaussian with mean and variance having been integrated, which can be evaluated by the closed form formula in [19]. The term $P(\{s_{abr}\}|\{s_{nmr}|z_{nm} = k, \backslash ab\}) = \prod_{r=1}^{R} P(s_{abr}|\{s_{nmr}|z_{nm} = k, \backslash ab\} \bigcup \{s_{abr'}\}_{r'=1..r-1})$. In this term, $P(s_{abr}|S)$ is equal to the proportion of elements in S whose values equal to s_{abr}. Other such terms are evaluated similarly. When sampling z_{ab}, to avoid underflow caused by the product of many probabilities in rule (3), we apply log to each term and then add a constant to each probability measure in log form for scaling to get a well-represented unnormalized distribution of z_{ab} for normalization.

In practice, sampling z,s,T alternatively from the very beginning does not yield good learning results as too many hidden variables are coupled. Instead, we firstly separate different users' data, and for each user degenerates the model to single-user model by setting $K = 1$ and then samples his hidden variables T and s using rule (2) and (4) until convergence. We then regard the currently assigned values for all s and T as initial values, reset K to some number bigger than 1, combine all users' data together and alternatively sample each user's z,s,T using (2),(3) and (4) until convergence.

Then we obtain a sampler which we can use to sample data from $P(\{z, s, T\}_{NMR}|D)$ to estimate the posterior expectation of each parameter given only the observed data. Specifically, we continuously sample sufficient $\{z, s, T\}_{NMR}$ from the conditional distribution $P(\{z, s, T\}_{NMR}|D)$, and then for each sample, evaluate $E_{\theta \sim P(\theta|\{z,s,T\}_{NMR},D)}[\theta]$ which has trivial closed form, and then compute their average for each parameter θ obtaining $E_{\theta \sim P(\theta|D)}[\theta]$ according to formula (1). This process amounts to firstly learning each user's individual behavior pattern separately, and then detecting behavior clustering structure by grouping similar behaviors together in a probabilistic way.

This learning algorithm is time and space efficient. In implementation, we use several matrices to store the empirical expectation of sufficient statistics of each local conditional distribution in the graphical model, and thus the resampling of a record takes $O(1)$ time to update the matrices. In this case, the learning algorithm takes linear time, that is, $O(NMR)$ time for each iteration, and the space complexity is proportional to the number of sufficient statistics, that is, $O(KLH)$, where K,L,H are typically set very small.

We now present how to make use of the learned behavior patterns to infer users' future activities. Generally, we would like to infer which cell tower is likely to be the one near a given user at some future time point t. Formally, this amounts to inferring the conditional distribution of x given t, namely $P(x|t)$. If we just use the behavior pattern learned from a single user's data to infer the activity of that user, by applying standard inference algorithm to the single-user model, we obtain the following result:

$$P(x|t) \propto \sum_s \Big(P(x|s)P(s) \sum_T P(t|s, T)P(T|s) \Big) \quad (5)$$

After learning, all probability terms in (5) can be computed using the learned parameters, thus $P(x|t)$ can be evaluated. In addition, the entropy of $P(x|t)$ can be used to assess the predictability of a user's location at a specific time point.

In addition to using the multi-user model, user behavior clustering can also be achieved by applying the KL-divergence to measure the similarity between any two users' individual behavior patterns $\{P(t|s)|s \in S\}$ that are learned from the single-user model as a distance function, and then use any well-studied clustering algorithm such as K-means to detect the behavior-clustering structure. The key point is that our definition of user daily behavior pattern as a set of temporal distributions makes the measurement of behavior similarity easy and informative with the support of information theory.

EXPERIMENTS AND EVALUATIONS

In this section, we conduct extensive experiments to show the effectiveness of our models in sensing individual behavior patterns, uncovering behavior similarity and differences among multiple users, and inferring user activities. For behavior pattern learning and behavior cluster detection, we compare the experiment results with the survey information associated with this data set for qualitative evaluation, while for user activity inference, we make quantitative evaluation

by using the extension of ROC graph in multi-class setting.

Single-user Behavior Learning

We firstly apply the single-user model to some selected users who have sufficient data recorded and learn model parameters for each of them separately using the algorithm presented. It should be noted that, although the framework is independent of the number of states, setting too many states may lead to over-fitting and hence make the learning results hard to interpret. For our scenario, we set 2 states(L=2) in the model and allow 2 temporal Gaussians in each state(H=2), which is sufficient to model users' routine behaviors. Using the learned parameters, we plot the behavior pattern $\{P(t|s)|s = 1, 2\}$ for some users in figure 3. For each user, the solid curve stands for $P(t|s = home)$ and the dashed curve stands for $P(t|s = work)$(may be unnormalized for ease of comparison), both of which have two Gaussian components, where "Home" and "Work" are the physical meanings we interpreted for the two states after learning. It can be seen that our single-user model has discovered very clear and meaningful daily behavior pattern for each of these users. User 8 apparently adopts a pattern of staying at home late till noon and then keeps working until midnight. In sharp contrast, both user 4 and user 27 stay at home in the morning and evening while working in the "middle", but user 4's working time shifts to the right compared with user 27, implying a later working schedule. In fact, we see three clear behavior patterns from these 6 users' individual curves, which are substantiated by the survey data saying that the working hours of users 17, 27 are 9am-5pm, of users 4, 81 are 11am-8pm, of users 8,23 are 2pm-11pm. It should be emphasized that such results are found by our unsupervised model automatically without requiring the state of each mobile record be labeled before learning, which thus is superior to [11]. Moreover, under our behavior representation, it is easy to compare the similarity of two users' behavior patterns by applying KL-divergence to two sets of distributions, which cannot be achieved using the behavior representation in [9]. For example,the KL-divergence between user 4 and user 81 is smaller than user 4 and user 27, implying a stronger similarity between the former pair of users.

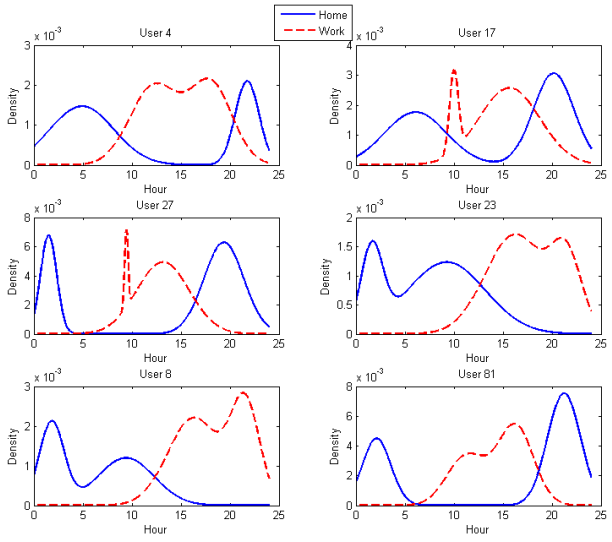

Figure 3. Behavior patterns of selected users learned by single-user model

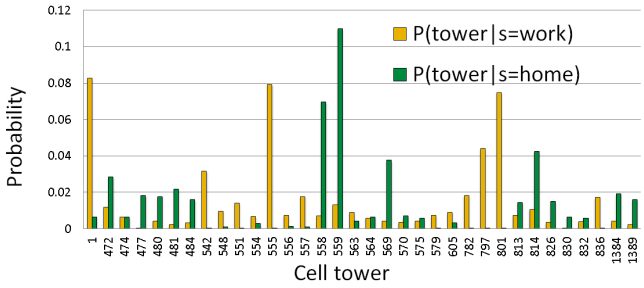

Figure 4. State-conditioned cell tower distribution for user 17

One byproduct of the learning process on single-user model is the cell tower distribution of each state, which is encoded in $P(x|s, \phi_n)$. We plot in figure 4 the cell tower distribution of both "work" state and "home" state learned for user 17. It is obvious that different states are associated with apparently different cell towers, e.g., his work state is mostly associated with cell tower 1, 555, 801, while his home state is primarily associated with cell towers 558, 559. In figure 4, the entropy of the cell tower distribution conditioned on work state(1.86) is relatively large compared with that on home state(1.67), which implies that this user's location is not very fixed at work state, suggesting a potentially varying working pattern.

Multi-user Behavior Learning

Though the single-user model can discover certain apparent patterns involving both temporal and spatial aspects in a user's daily behavior, learning each user's data separately might suffer from over-fitting a single user's sparse data, and thus cannot well disclose the most typical characteristic behavior patterns hidden in a group of users. For example, the temporal distribution on work state of user 27 plotted in figure 3 has a severe singularity, probably due to data sparsity in neighboring time intervals of this user. Also, modeling a single user's data separately implicitly assumes that the user's behavior remains relatively fixed over a long time. Although this might be true for many users who lead highly regular lives, there exists some users who adopt varying life styles, which cannot be well captured by single-user model. To solve these problems, we carry out experiments applying multi-user model to 30 selected users' data as a whole.

We set K, the number of typical behavior patterns, to 4 in the model setting. We plot the state-conditioned temporal distributions of the top 3 learned behavior patterns in figure 5. These are not some users' individual behavior patterns, but rather the typical latent behavior patterns learned by exploiting all users' data in a collaborative way. Each of them represents a common behavior pattern shared by a group of users who behave similarly, and the differences between them are reflective of the major difference between different group of users. These three patterns learned by our multi-user model happen to reflect three kinds of typi-

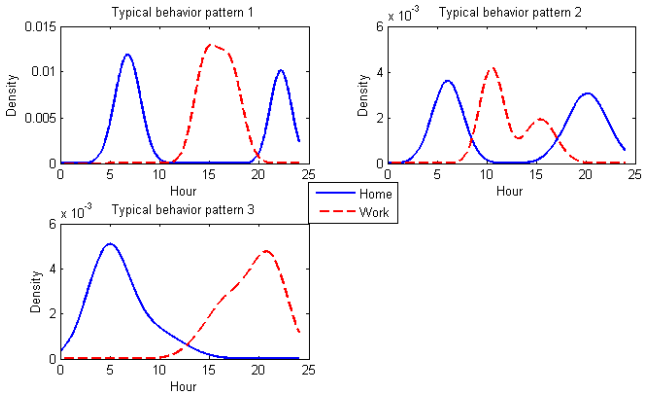

Figure 5. Top 3 typical behavior patterns learned by multi-user model

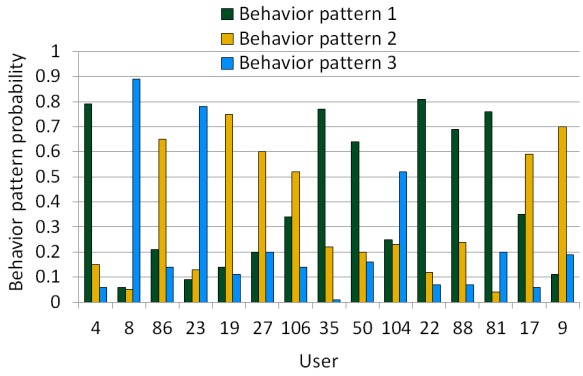

Figure 6. Behavior mixture parameters(γ) of selected users

cal working hour schedules in the survey data, namely 9am-5pm(pattern 2), which characterize most staff members and Business school students, 11am-8pm(pattern 1) and 2pm-11pm(pattern 3), which most likely characterize two typical life patterns of students in Media Lab. This result demonstrates the ability of our multi-user model to separate informative behavior patterns mixed in multiple users' data, thus giving us a clue about the underlying social structure. Also, note that the curves in figure 5 are more smooth than those in individual behaviors plotted in figure 3, implying that the multi-user model well overcomes the sparsity problem in single user's data through appropriate sharing of information among users with similar behavior patterns.

The multi-user model also learns for each user a γ parameter, indicating at what probability this user adopts each of those K behaviors in his life. We plot some selected users' behavior mixture parameters(γ) in figure 6 for comparison. We can clearly see that user 4 leads a very regular life, which is dominated by behavior pattern 1. Similar things happen to user 8 but with a different dominant behavior, while user 104 adopts a relatively varying life style, a fact which cannot be captured by the single-user model for the reasons discussed before. The behavior characterization of 79% users by the multi-user model are approximately consistent with the personal working schedule information in the survey data.

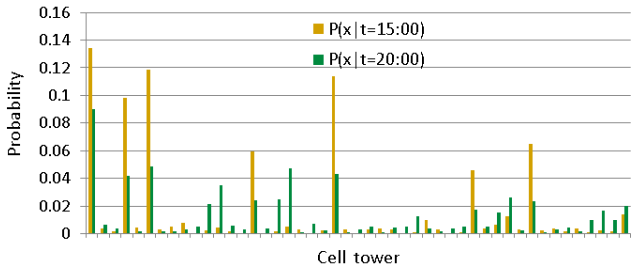

Figure 7. Time-conditioned cell tower distribution for user 4

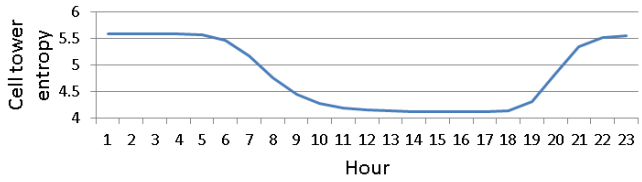

Figure 8. Daily change of cell tower entropy for user 4

User Activity Inference

Next we experiment the inference of user activity, i.e., infer the possible location of a user given a specific time using the behavior pattern learned from his historical data. This is done by applying the conditional distribution derived in (5), whose needed parameters are learned from the target user's historical data using single-user model. For an intuitive understanding of how time determines a user's location, we select user 4, and for each hour t plot his conditional cell tower distribution $P(x|t)$ which is evaluated using (5) with the parameters learned for him. $P(x|t = 15)$ and $P(x|t = 20)$ are shown in figure 7 with the tower id's erased. It can be seen that at work time(15:00), the tower distribution has a low entropy, implying high predictability of the user's location, while at 20:00, which is the time for the user to transit between two states, his location is much less predictable as evidenced by the high entropy of the distribution at this time. This fact can be seen more clearly in figure 8, which shows how the entropy of user 4's cell tower distribution varies over t in a day, where we can see that the uncertainty of cell tower remains the lowest at work state, and increases sharply when the user transits from work to home.

For quantitative inference, we prepare for each selected user a held-out test set containing randomly selected 5 days' data. Each record in the test set contains time $t^\star$ which is the input to the predictor, and the cell tower $x^\star$ where the user is currently located, which is the ground truth used for evaluation. Then, for each $t^\star$, we infer cell tower x by evaluating (5) using the learned parameters for this user, obtaining $P(x|t^\star)$, which is essentially a probability vector $\mathbf{p}$ satisfying the property $\sum_{x=1}^{F} p_x = 1$ with p_x indicating the inferred probability that this user is near tower x. We note that this is essentially a multi-class classification problem, with each cell tower representing a class. To evaluate the inference performance objectively, we make use of the extension of ROC curve in multi-class setting [13]. Specifically, the area under ROC curve(AUC) is a standard metric

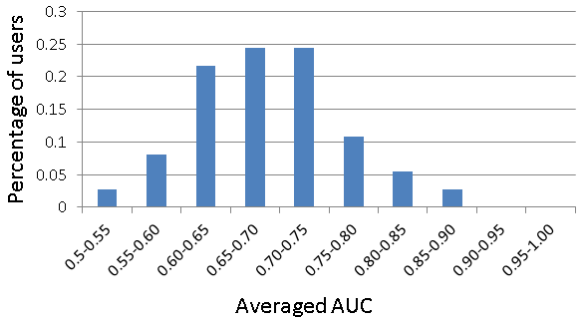

Figure 9. Histogram of the averaged AUC of activity inference on 37 selected users

to measure the performance of a binary classifier by evaluating to what extent the classifier ranks a positive instance higher than a negative one [12]. To extend it to multi-class setting, note that multi-class classification can be cast to a set of pair-wise binary classification problems, and we estimate the AUC for each such binary classifier using the corresponding pair of class probabilities in the prediction vector $P(x|t)$, and then compute their average. We conducted experiments on 37 selected users and for each of them estimate his average AUC. The distribution plotted in figure 9 shows that a certain number of users yield an average AUC larger than 70%, demonstrating our model's ability in capturing the structures of users' routine behaviors and inferring users' activities. As we target at modeling users with regular behaviors, those leading highly entropic lives cannot be well described by our model, as shown by the relatively low AUC in figure 9. This diversity illustrates our model's ability in distinguishing users by their behavior regularity.

CONCLUSION AND FUTURE WORK

In this paper, we have developed techniques to automatically understand human behaviors from mass amount of mobile data. To that end, we firstly define a probabilistic representation for users' regular behavior patterns, which differs from prior work in that it captures the most essential and informative features as well as encodes uncertainty in the internal structures of human's regular behavior patterns, can be easily exploited to measure the behavior similarity between multiple users and infer users' activities. We then propose a hierarchical Bayesian network which can learn each user's characteristics behavior pattern, uncover behavior clustering structure through an appropriate information sharing among different users' data, and capture users' behavior variation. Experiments performed on a public reality mining data set with rich survey data as ground truth demonstrate the effectiveness of our models in both understanding human behaviors and inferring human activities.

One limitation of the model lies in its inability to identify periodic variation of humans' routine behaviors, such as adopting one routine on weekdays and another on weekends. Although we can capture such periodic variations by separating the data according to domain knowledge and then applying the model to each data partition independently, how to make this process automatic remains an open problem. Also, we attempt to relax our parametric model to a non-parametric one which can automatically determine the number of a user's typical states and the number of typical behavior patterns in a group of users. We also plan to address how to effectively uncover user's behavior patterns from more sparse mobile data, as in traditional mobile phone data set.

ACKNOWLEDGEMENT

This research was supported in part by Hong Kong, Macao and Taiwan Science & Technology Cooperation Program of China under Grant No. 2012DFH10010. We thank Nathan Eagle for his kind help on the Reality Mining data set.

REFERENCES

1. Reality Mining Project. `http://reality.media.mit.edu`.
2. J. Begole, J. Tang, and R. Hill. Rhythm modeling, visualizations and applications. In *Proc. User interface software and technology 2003*, pages 11–20. ACM, 2003.
3. C. Bishop. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.
4. D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Machine Learning*, 3:993–1022, 2003.
5. T. Choudhury and A. Pentland. Sensing and modeling human networks using the sociometer. In *Proc. ISWC 2003*, volume 1530, 2003.
6. D. Choujaa and N. Dulay. Predicting human behaviour from selected mobile phone data points. In *Proc. UbiComp 2010*, pages 105–108. ACM, 2010.
7. B. Clarkson. *Life Patterns: structure from wearable sensors*. PhD thesis, Massachusetts Institute of Technology, 2002.
8. N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.
9. N. Eagle and A. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.
10. N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.
11. K. Farrahi and D. Gatica-Perez. What did you do today?: discovering daily routines from large-scale mobile data. In *Proc. AMMM 2008*, pages 849–852. ACM, 2008.
12. T. Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine Learning*, 31:1–38, 2004.
13. D. Hand and R. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.
14. M. Kim and D. Kotz. Periodic properties of user mobility and access-point popularity. *Personal and Ubiquitous Computing*, 11(6):465–479, 2007.
15. D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.
16. L. Liao. Learning and inferring transportation routines. *Artificial Intelligence*, 171(5):311–331, 2007.
17. A. Noulas, M. Musolesi, M. Pontil, and C. Mascolo. Inferring interests from mobility and social interactions. In *Workshop on Analyzing Networks and Learning with Graphs*, 2009.
18. D. Peebles, H. Lu, N. Lane, T. Choudhury, and A. Campbell. Community-guided learning: Exploiting mobile sensor users to model human behavior. In *Proc. AAAI 2010*, 2010.
19. Y. W. Teh. The Normal Exponential Family with Normal-Inverse-Gamma Prior. `http://www.gatsby.ucl.ac.uk/~ywteh/research/notes/GaussianInverseGamma.pdf`.
20. Y. Zheng, Q. Li, Y. Chen, X. Xie, and W. Ma. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 312–321. ACM, 2008.