Discovering Routines from Large-Scale Human Locations using Probabilistic Topic Models

KATAYOUN FARRAHI and DANIEL GATICA-PEREZ
IDIAP Research Institute
Ecole Polytechnique Fédérale de Lausanne (EPFL)

3

In this work, we discover the daily location-driven routines that are contained in a massive real-life human dataset collected by mobile phones. Our goal is the discovery and analysis of human routines that characterize both individual and group behaviors in terms of location patterns. We develop an unsupervised methodology based on two differing probabilistic topic models and apply them to the daily life of 97 mobile phone users over a 16-month period to achieve these goals. Topic models are probabilistic generative models for documents that identify the latent structure that underlies a set of words. Routines dominating the entire group's activities, identified with a methodology based on the Latent Dirichlet Allocation topic model, include "going to work late", "going home early", "working nonstop" and "having no reception (phone off)" at different times over varying time-intervals. We also detect routines which are characteristic of users, with a methodology based on the Author-Topic model. With the routines discovered, and the two methods of characterizing days and users, we can then perform various tasks. We use the routines discovered to determine behavioral patterns of users and groups of users. For example, we can find individuals that display specific daily routines, such as "going to work early" or "turning off the mobile (or having no reception) in the evenings". We are also able to characterize daily patterns by determining the topic structure of days in addition to determining whether certain routines occur dominantly on weekends or weekdays. Furthermore, the routines discovered can be used to rank users or find subgroups of users who display certain routines. We can also characterize users based on their entropy. We compare our method to one based on clustering using K-means. Finally, we analyze an individual's routines over time to determine regions with high variations, which may correspond to specific events.

Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems; I.2.6 [**Artificial Intelligence**]: Learning; I.5.4 [**Pattern Recognition**]: Applications

General Terms: Human Factors, Algorithms

Additional Key Words and Phrases: Human activity modeling, topic models, reality mining

This research was supported by the Swiss National Science Foundation through the MULTI project.
Authors' address: IDIAP Research Institute, Martigny, Switzerland and Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland; email: kfarrahi@idiap.ch.

DOI 10.1145/1889681.1889684 http://doi.acm.org/10.1145/1889681.1889684

ACM Reference Format:
Farrahi, K. and Gatica-Perez, D. 2011. Discovering routines from large-scale human locations using probabilistic topic models. ACM Trans. Intell. Syst. Technol. 2, 1, Article 3 (January 2011), 27 pages.
DOI = 10.1145/1889681.1889684 http://doi.acm.org/10.1145/1889681.1889684

1. INTRODUCTION

The mobile phone is the most widely spread technology, with more than half of the human population (4 billion) having subscriptions in the world by 2008 [Ling and Donner 2009]. This revolution in mobile communication is very recent, and the progression occurred rapidly, simply over the last thirty years. Recently, researchers are realizing the rich information that can be captured by these ubiquitous devices, which has potential impact on a vast range of domains including epidemiology, psychology and sociology, urban planning, security and intelligence, and health monitoring. The rich and vast amount of data that can be captured by these ubiquitous devices ranges from communication, location, proximity, motion, and video, to name a few. The potential for use of this information is huge and mostly unexplored. The potential for mobile phones to help humans in aspects of life other than communication in the future is huge, and what the next mobile revolution will lead to is clearly open.

Our focus is the automatic discovery of human activities and routines from mobile phone location data collected by one hundred individuals over the course of a year. We define routines to be temporal regularities in people's lives. A routine often involves patterns of location transitions over time (e.g., being at work or going from work to home), possibly over different time scales and for varying time intervals. Automatic routine classification and discovery are in general challenging tasks as people's locations often vary from day to day and from individual to individual, and data from sensors can frequently be incomplete as well as noisy. A supervised learning approach to activity recognition would require prior knowledge in the form of predefined activity categories and labeled data [Liao et al. 2006]. In contrast, an unsupervised learning approach has the potential of automatic discovery of emerging routines of people not requiring training data. Through discovery, sifting through large amounts of noisy data becomes possible. Further, one can cluster data (i.e., people or days) corresponding to the most common routines (those of several people) and discover the dataset structure with minimal prior knowledge.

In this work, we develop a novel methodology built on topic models to discover location-driven routines; these models were initially designed for text documents [Blei et al. 2003; Rosen-Zvi et al. 2004]. Recently, they have been successfully applied to data sources other than text, such as images [Monay and Gatica-Perez 2007; Quelhas et al. 2007], video [Niebles et al. 2006], genetics [Pritchard et al. 2000], and wearable sensor data [Huynh et al. 2008], but to our knowledge, their use for real-life routine modeling from large-scale mobile phone data is novel. Topic models are generative models that represent documents as mixtures of topics, learned in a latent space, and they allow for clustering and ranking of documents, words, and other entities, like

authors [Rosen-Zvi et al. 2004]. They are advantageous to activity modeling tasks due to their ability to effectively characterize discrete data represented by bags (i.e., histograms of discrete items). These models can capture which words are important to a topic as well as the prevalence of those topics within a document, resulting in a rank measure. The fact that multiple topics can be responsible for the words occurring in a single document discriminates this model from standard Bayesian classifiers [Duda et al. 2000]. We can take advantage of the bag flexibility to find routines at different temporal granularities, additionally incorporating transitions over time. In this article, we show that topic models prove to be effective in making sense of behavioral patterns at large-scale while filtering out the immense amount of noise in real-life data.

The contributions of this work are the following.

(1) We devise a novel bag representation of a day of the life of a mobile phone user which captures both fine-grain and coarse-grain times as well as transitions in locations over time.
(2) We propose a methodology for the automatic discovery of daily location-based routine patterns with Latent Dirichlet Allocation (LDA) [Blei et al. 2003], where we discover routines characteristic of all days in the dataset.
(3) We extend our methodology via the Author Topic model (ATM) [Rosen-Zvi et al. 2004] to discover routines of a varying sort, this time emphasized on individual users' as opposed to all users' days.
(4) We perform several analysis tasks with the model outputs, including finding routines which dominate on certain types of days; finding days which are well represented by few/many topics; finding a given user's dominating daily patterns; finding low entropy and high entropy users; determining when a large variation occurs for a given user's routine over time; and discovering groups of users that follow certain trends.

This article is organized as follows. The next section discusses related work. Section 3 outlines the overview of this research and our approach. We then discuss our bag representation methodology in Section 4, followed by a brief overview of topic models in Section 5. We present and discuss the experimental results in Section 6. Section 7 draws the article to conclusion.

2. RELATED WORK

There is relatively little work on activity recognition and modeling tasks using mobile phone data. Research using mobile phone data has mostly focused on location-driven data analysis, more specifically, using Global Positioning System (GPS) data to predict transportation mode [Patterson et al. 2003, Reddy et al. 2008], to predict user destinations [Krumm and Horvitz 2006] or paths [Akoush and Sameh 2007], and to predict daily step count [Sohn et al. 2006]. Other location-driven tasks have made use of Global System for Mobile Communications (GSM) data for indoor localization [Otsason et al. 2005] or WiFi for large-scale localization [Letchner et al. 2005]. The BeaconPrint algorithm [Hightower et al. 2005] uses both WiFi and GSM to learn the places a user goes and detect if the user returns to these places.

On the other hand, there is an increasing body of work on activity recognition using various types of wearable sensor data [Choudhury et al. 2006, Tapia et al. 2007] not including mobile phones. For example, the sociometer [Choudhury and Pentland 2003] is a wearable sensor package, which is used to monitor face-to-face interactions and social dynamics. PlaceLab [Larson and Intille] is an example of a "living lab", where hundreds of sensors are built into objects and the home environment (as opposed to wearable sensors) for various research purposes including activity recognition [Intille et al. 2006, Logan et al. 2007, Tapia et al. 2004]. Similar work has been done in an office space in which hundreds of motion sensors were used to study the social interactions and behaviors of approximately 100 subjects [Wren et al. 2007]. In Liao et al. [2006], GPS data from wearable sensors is used for place labeling, specifically, recognizing significant locations and associating activities to these locations, such as "walking", "visiting", and "leisure".

Two works related to human activity modeling from mobile sensor data are Eagle and Pentland [2009] and Gonzalez et al. [2008]. The first of these works [Eagle and Pentland 2009], uses principle component analysis (PCA) to identify the main components which structure daily human behavior from the Reality Mining dataset [Eagle et al. 2009]. These main components are a set of characteristic vectors, termed *eigenbehaviors*. To define the daily life of an individual in terms of eigenbehaviors, the top eigenbehaviors will show the main routines in the life of a group of users, for example, being at home overnight. The role of the remaining eigenbehaviors is to describe the more precise, nontypical behaviors in individuals' or the group's lives. Eigenbehaviors are described over what we consider to be fine-grain locations (30-minute time intervals) and are representative of the entire day's activities as opposed to morning only or evening only. Results are presented for location data only where Bluetooth and raw location data is used in an HMM structure to infer a user's location for a given time. In our work, we present a methodology based on a bag of location sequences structure that is advantageous over Eagle and Pentland [2009] in that it contains both fine-grain and coarse-grain time considerations, which keeps into account transitions in location and is robust to variations in the data that may be due to noise or due to variations in the dataset, such as eating lunch at 11:30 am as opposed to 11:55 am. Further, due to our location sequence structure, we can discover routines characteristic of various intervals in the day. Our topic model methodology clearly defines a mechanism to rank users and days (with probabilities always greater than zero unlike Eagle's methodology), and with easily identifiable routines with semantic meanings which can be visualized comprehensibly over many of the discovered topics. Ranking allows us to see the raw data in a particular order (given by probabilities), giving structure to the data. This is true for both users and days and is useful for visualizing and structuring the data. We can perform several tasks with the discovered data, such as find users that go to work early, find groups of users that are at home during the day, or find users that turn their phones off in the morning. We also discover a varying sort of routines based on the ATM, a differing methodology which incorporates individual user identities into the model. With this, we can rank users and days, characterize users and day structures based on the

number of topics composing most of the probability mass, and analyze an individual's daily life patterns over time. Finally, PCA is a traditional technique in pattern recognition [Duda et al. 2000]. The more modern techniques we are considering are state-of-the-art models and an active domain in machine learning.

The second work related to activity modeling from mobile sensor data is by Gonzalez et al [2008]. Recently, they used mobile phone data to study the trajectories of human mobility patterns, and found that human trajectories show a high degree of temporal and spatial regularity, more specifically, that individual travel patterns can collapse into a single spatial probability distribution showing that humans follow simple, reproducible patterns. This study was performed on a large-scale mobile phone dataset of over 100, 000 users over a period of six months.

Another closely related work published simultaneously to ours [Farrahi and Gatica-Perez 2008a, 2008b], is described in Huynh et al. [2008] and uses topic models for human activity discovery, using wearable sensor data and not mobile phone data. The method identifies activity patterns in one single person's daily life over sixteen days, using two wearable sensors, one placed on the right hip and the other on the right wrist. For activity recognition, the Latent Dirichlet Allocation (LDA) topic model is used where activity words are manually labeled or automatically recognized low-level activities, and the topic model is used to discover patterns in these activities, which are essentially co-occurring low-level activities. Further, a document is constructed from a sliding window of length D. In contrast, our work investigates the human routine discovery task from mobile phone data, on a large scale, and we use this data to discover group routines in addition to individual routines. Our documents are independent and identically distributed as in topic models, though this is not the case in Huynh et al. [2008]. Further, our methodology has proven successful on lower level input data which can be obtained more directly from sensor data, such as the locations of an individual and their proximate interactions [Farrahi and Gatica-Perez [2008a]]. The methodology proposed in Huynh et al. [2008] requires higher level information regarding a person's activities through the use of multiple devices attached to various body parts, unlike our work which only relies on one single device (a phone) which is worn and used naturally. Further, our methodology also investigates the Author-Topic model [Rosen-Zvi et al. 2004].

A preliminary version of our work was published in Farrahi and Gatica-Perez [2008b], where we investigated the location-driven routines of a selection of 30 users over a four month period. Previously, we also studied the proximity dataset in addition to the location data using a conceptually simpler topic model [Farrahi and Gatica-Perez 2008a]. This work significantly extends our initial work by applying our methodology on the full dataset of 97 users and 16 months (491 days) of available data. The issue of model selection is also investigated, and the analysis is thorough and extensive considering both individual and group behavior. As discussed in the introduction, we also perform several new tasks with the routines discovered.

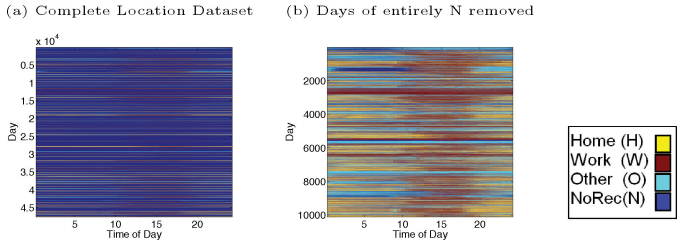


Fig. 1. Visualizations of the location data for (a) all the users and the entire set of days and (b) all the users and days excluding days which contain entirely no data. The x axis corresponds to the time of day (in hours). The y axis corresponds to days.

3. OVERVIEW OF OUR WORK

We use the Reality Mining dataset [Eagle et al. 2009], which contains the mobile phone sensor data recordings of 97 subjects studying or working at MIT over the 2004-2005 academic year. Reality Mining has been recently named by MIT Technology Review magazine as "one of the 10 technologies most likely to change the way we live" [MIT Technology Review]. The dataset contains the location (cell tower connections), proximity (Bluetooth connections), communication as well as phone application usage of the subjects, though much of this data is noisy and missing. Here we focus on the location dataset, which is given by cell tower connections. Throughout the study over 32, 000 towers were recorded, to which we assign semantic labels. We assign 'location labels' of *home* (H), *work* (W), or *other (or out)* (O) to the towers using labels provided by the collectors of the dataset. More precise details regarding location labeling and the dataset can be found in Sections 4.1 and 6.1, respectively. At this point, we simply assume that we can represent a day in the life of a mobile phone user in terms of location labels for visualization and description purposes. Assuming we can express the day in the life of a person's locations in terms of these labels, with the addition of a fourth *no reception* (N) label in case the phone was off or no data was recorded, we can then visualize the users' location patterns as a function of time of day, as in Figures 1(a) and 1(b). Each row in the figures is a day of a person's life in terms of his/her location, where the x-axis is the time of day, and the four colors represent the four location labels. Figure 1(a) shows our entire dataset for the 97 users and their 491 days of activities, many of which contain *no reception* the entire day. Figure 1(b) shows the input dataset used in which we remove days containing entirely *no reception* labels. Looking at Figure 1(b), the immense quantity of noise and missing data becomes apparent as well as the amount of data and complex mixture of activities which potentially exist. In addition, it is not apparent how to determine dominating group routines and how to characterize individuals in terms of the groups' routines. These are a few of the points we address with our proposed methodology, illustrated in Figure 2.

Our overall goal is to determine what human routines are contained in mobile tower connection data and how to discover them in an unsupervised manner. As described earlier, we represent a day in the life of an individual in terms of their locations obtained by cell tower connections and use this

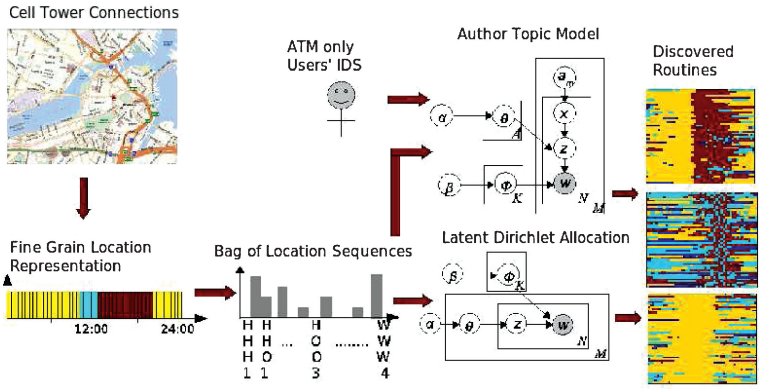


Fig. 2. Simplified block diagram of our methodology. User locations, given by cell tower connections, are transformed into bag of location sequences by first representing a user location as a fine grain location representation. This bag of location sequences is passed to Latent Dirichlet Allocation, in the first set of experiments, resulting in the discovery of routines. In the second set of experiments, the bag of location sequences and the user IDs are input to the Author-Topic Model, resulting in the discovery of routines of a varying sort.

information to form a *bag of location sequences*. This bag representation was carefully designed to capture dynamics (i.e., location transitions) as well as both fine-grain (30 minute) and coarse-grain (several hour) time considerations. Details of the method for bag construction are in Section 4. Overall, we make an analogy between the bag of location sequences (or words) for mobile data and a bag of words for text documents, where a location sequence is analogous to a text word, a day in the life of a person is analogous to a document, and a person is analogous to the author of a document. We use two models to discover routines. The first is Latent Dirichlet Allocation (LDA), illustrated in Figure 2, in which the input is the bag of location sequences. The second is based on the Author Topic Model (ATM), also visualized in Figure 2, in which user identity is input to the model in addition to the location sequences. The output of the models is a set of probability distributions over words and latent topics, capturing the dominating underlying routines in the dataset. We can then rank location sequences and days per topic, as well as users per topic in the case of ATM, and observe the routines discovered as topics.

4. BAG REPRESENTATIONS

In this work, we design a bag representation for location sequences. Location sequences are not suitable for topic models in their original time sequence form since words in the topic model should be interchangeable. By constructing a bag representation to capture fine- and coarse-grain location, both can be encoded and can be viewed as analogous to words for text mining.

4.1 Fine-Grain Location Representation

For a given individual in the dataset, there are entries for cell towers users connect to, the start and end connection times. Over 32 000 towers are seen

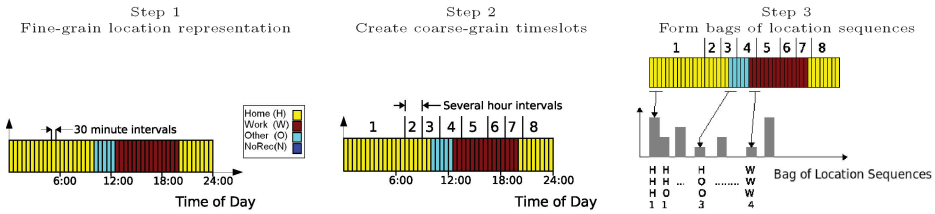


Fig. 3. The construction of the bag of location sequences explained illustratively in a 3-step process. Step 1 is the division of a day into fine-grain 30-minute time intervals, which we call the fine-grain location representation. This first step includes associating a single location label of H, W, O, or N to each 30-minute time interval. Step 2 is division of a day into 8 unevenly distributed coarse-grain time intervals. Step 3 demonstrates the word construction. Three consecutive fine-grain location labels and the coarse-grain time interval are combined to form location sequences. The set of location sequences for a day forms the bag of location sequences.

by all the users phones over the course of the year. Since we are interested to discover routines existing in the location data, we classify the cell towers into 3 semantic categories, removing the noise of the actual tower ID. As stated in Section 3, the categories were *home* (H), *work* (W), and *other* (O), representing towers which were the self-declared homes of users, work towers at MIT campus, and other towers, respectively. An initial set of labels was obtained from the Reality Mining creators. The list of W towers obtained from MIT was incomplete as several students never connected to any of those towers and thus were never considered to be at work. To resolve this issue, additional W labels were inferred from being in proximity to each person's computer; we did not consider being in proximity to one's laptop as being at work due to the mobile nature of the device. There was a fourth *no reception* (N) label, applied when there was no tower connection recorded for a user at a given time.

Following the labelling of cell towers into location categories, the day in the life of a user can then be expressed as a sequence of these location labels. The first step in forming our bag of location sequences is the construction of a fine-grain location representation, as illustrated in Figure 3, Step 1. We chose to a divide a day into fine-grain, 30-minute time intervals, resulting in 48 time blocks per day. We use 30-minute slots as many events of daily life are synchronized with half-hourly schedules and this does not result in vocabulary size explosion as discussed in the following section. For each block of time, we choose the location label which occurred for the longest duration, resulting in a single location label per timeslot. This is an important step as tower connections can be noisy and fluctuative. The result is a day of a user represented as a sequence of 48 location labels, visualized for the entire input dataset in Figure 1(b).

4.2 Bag of Location Sequences

The bag of location sequences is built from the fine-grain location representation considering 8 coarse-grain timeslots in a day, as shown in Figure 3, Step 2. We divide a day into the timeslots as follows: 0-7 am (1), 7-9 am (2), 9-11 am (3), 11 am-2 pm (4), 2-5 pm (5), 5-7 pm (6), 7-9 pm (7), and 9-12 pm (8). The goal

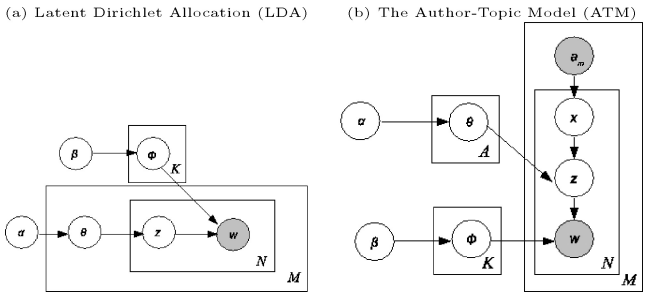


Fig. 4. Graphical models of two probabilistic topic models (a) Latent Dirichlet Allocation (LDA) and (b) the Author Topic Model (ATM).

of these coarse-grain timeslots is to remove some of the potential noise due to minor time differences between daily routines. For example, if a user leaves the house at 7:30 am as opposed to 8am, we want to capture the important feature of "leaving the house early in the morning" and not the minor time difference of this routine between days. The choice of the timeslots is also guided by common sense about daily activities (e.g., typical lunch times, working times, sleeping times).

Finally, the third step in building the bag of location sequences is the word construction, visualized in Figure 3, Step 3. A location sequence contains 3 consecutive location labels in the fine-grain representation, corresponding to 1.5 hour intervals, followed by one of the 8 timeslots in which it occurred. Thus a location sequence has 4 components, 3 location labels followed by a timeslot. We take overlapping 1.5-hour sets of labels to make a location sequence, so that if we had a pattern HHHOW in the interval 7 am-9:30 am, we would have for 7:30 am, 8 am, and 8:30 am, the following location sequences: HHH1, HHO1, and HOW1, where 1 indicates timeslot 1. Finally, the bag of location sequences is the histogram of the location sequences present in the day. In this article, a document is a day of a user and an author is an individual.

5. TOPIC MODELS FOR ROUTINE DISCOVERY

Topic models are powerful tools initially developed to characterize text documents, but can be extended to other collections of discrete data. They are probabilistic generative models that can be used to explain multinomial observations by unsupervised learning. Formally, the entity termed *word* is the basic unit of discrete data defined to be an item from a vocabulary of size V. A *document* is a sequence of N words. A *corpus* is a collection of M documents. There are K latent topics in the model, where K is defined by the user.

5.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Figure 4(a)) is a generative model, introduced by Blei et al. [2003], in which each document is modeled as a multinomial distribution of topics and each topic is modeled as a multinomial distribution of words. The generative process begins by choosing a distribution over topics $\mathbf{z} = (z_{1:K})$ for a given document. Given a distribution of topics for a document,

words are generated by sampling topics from this distribution. The result is a vector of N words $\mathbf{w} = (w_{1:N})$ for a document.

LDA assumes a Dirichlet prior distribution on the topic mixture parameters θ and ϕ, to provide a complete generative model for documents. θ is an M x K matrix of document-specific mixture weights for the K topics, each drawn from a Dirichlet(α) prior, with hyperparameter α. ϕ is an V x K matrix of word-specific mixture weights over V vocabulary items for the K topics, each drawn from a Dirichlet(β) prior, with hyperparameter β.

The main objectives of LDA inference are to

(1) find the probability of a word given each topic k, $p(w = t|z = k) = \phi_k^t$, and
(2) find the probability of a topic given each document m, $p(z = k|d = m) = \theta_m^k$.

Several approximation techniques have been developed for inference and learning in the LDA model [Blei et al. 2003; Griffiths and Steyvers 2004]. In this work we adopt the Gibbs sampling approach [Griffiths and Steyvers 2004].

For the LDA model visualized in Figure 4(a), the following distributions hold:

$$p(\theta|\alpha) = p(\theta) \sim Dirichlet(\alpha) \quad (1)$$

$$p(\phi|\beta) = p(\phi) \sim Dirichlet(\beta) \quad (2)$$

$$p(z|\theta^{(d)}) \sim Multinomial(\theta^{(d)}) \quad (3)$$

$$p(w|z, \phi^{(z)}) \sim Multinomial(\phi^{(z)}) \quad (4)$$

where $\phi^{(z)}$ represents the word distribution for topic z, and $\theta^{(d)}$ represents the topic distribution for document d.

From the assumptions in equations (1)-(4), we obtain $p(w|z, \phi) = \prod_{k=1}^{K} \prod_{t=1}^{V} (\phi_k^t)^{n_k^t}$ where n_k^t is the number of times word t is assigned to topic k. n_k^t is also called the word-topic count and $n_k = \sum_{t=1}^{V} n_k^t$ is called the word-topic sum. We also obtain $p(z|\theta) = \prod_{m=1}^{M} \prod_{k=1}^{K} (\theta_m^k)^{n_m^k}$ where n_m^k is the number of times topic k occurs in document m. n_m^k is also called the topic-document count, and $n_m = \sum_{k=1}^{K} n_m^k$ is called the topic-document sum.

Further details of the Gibbs sampling for LDA model parameter estimation can be found in Griffiths and Steyvers [2004]. In practice, we can use the procedure summarized in Figure 5 to estimate the model parameters.

5.2 Author-Topic Model

The Author-Topic model (ATM), introduced by Rosen-Zvi et al. [2004] is also a generative model for documents that extends LDA to include authorship information. In ATM, each author is associated with a multinomial distribution over topics and each topic, like LDA, is associated with a multinomial distribution over words. By modeling the interests of authors, it becomes possible to establish what topics an author writes about, which authors are likely to have written documents similar to an observed document, and which authors produce similar work.

For ATM, each word in a document is associated with two latent variables, an author, x, and a topic, z. The graphical model in Figure 4(b) illustrates the process. The set of authors of document m is defined as a_m, where $A = |a_m|$ is the

// GOAL: Given a training corpus, α, β, and K, estimate the parameters n_m^k and n_k^t from which we can determine the model parameters $\hat{\phi}_k^t$ and $\hat{\theta}_m^k$.

// Initialization
1) Initialize the count parameters, $n_m^k = 0$, $n_k^t = 0$.
2) **Iterate** over each word w in the corpus:
  3) Sample a topic k from $k \sim Mult(\frac{1}{K})$.
  4) Update the count parameters n_m^k, n_k^t as follows $n_m^k = n_m^k + 1$, $n_k^t = n_k^t + 1$.

// Run the chain
5) **Iterate** over a large number of iterations (e.g. 1000):
  6) **Iterate** over each word w in the corpus:
    7) Decrement the current word w and current word's topic assignment t counts as follows $n_m^k = n_m^k - 1$, $n_k^t = n_k^t - 1$.
    8) Sample a topic k from $p(z = k|\mathbf{z}_{\neg i}, \mathbf{w}) \propto \frac{n_k^t+\beta}{\sum_{t=1}^V n_k^t+\beta} \cdot \frac{n_m^k+\alpha}{\sum_{k=1}^K n_m^k+\alpha}$.
    9) Increment the new word/topic and topic/document counts as follows $n_m^k = n_m^k + 1$, $n_k^t = n_k^t + 1$.

// Compute model parameters
10) Estimate the unknown parameters as follows
$\hat{\phi}_k^t = \frac{n_k^t + \beta}{n_k + V\beta}$, and $\hat{\theta}_m^k = \frac{n_m^k + \alpha}{n_m + K\alpha}$, where $\hat{\phi}$ and $\hat{\theta}$ are the model parameter estimates, $n_k = \sum_{t=1}^V n_k^t$, and $n_m = \sum_{k=1}^K n_m^k$.

Fig. 5. Gibbs sampling algorithm for LDA.

number of authors who generated the documents in the corpus. Furthermore, x indicates the author responsible for a given word, chosen from a_m. In this model, ϕ denotes the V x K matrix of word-topic distributions, with a multinomial distribution over V vocabulary items for each of K topics drawn independently from a Dirichlet(β) prior. θ is the A x K matrix of author specific mixture weights for these K topics, each drawn from a Dirichlet(α) prior.

The main objectives of ATM inference are to

(1) find the probability of generating word t from topic k, ϕ_k^t and
(2) find the probability of assigning topic k to a word generated by author a, θ_a^k.

For the ATM model visualized in Figure 4(b), the following distributions hold:

$$p(\theta|\alpha) = p(\theta) \sim Dirichlet(\alpha) \quad (5)$$
$$p(\phi|\beta) = p(\phi) \sim Dirichlet(\beta) \quad (6)$$
$$p(z|x, \theta^{(x)}) \sim Multinomial(\theta^{(x)}) \quad (7)$$
$$p(w|z, \phi^{(z)}) \sim Multinomial(\phi^{(z)}) \quad (8)$$
$$p(x|a_m) \sim Uniform(a_m) \quad (9)$$

where $\theta^{(x)}$ represents the topic distribution for authors x.

For Gibbs sampling, the joint conditional probability distribution defined in Step 9 of Figure 6 is used [Rosen-Zvi et al. 2004], where the word-topic count, n_k^t is the number of times word t is assigned to topic k and $n_k = \{n_k^t\}_{t=1}^V$ is the word-topic sum. The topic-author count, n_a^k, is the number of times author a

// GOAL: Given a training corpus, α, β, and K, estimate the parameters n_a^k and n_k^t from which we can determine the model parameters $\hat{\phi}_k^t$ and $\hat{\theta}_a^k$.

// Initialization
1) Initialize the count parameters, $n_a^k = 0$, $n_k^t = 0$.
2) **Iterate** over each word w in the corpus:
  3) Sample a topic k from $k \sim Mult(\frac{1}{K})$.
  4) Sample an author a from $a \sim Mult(\frac{1}{A_m})$ where A_m is the list of authors of document m.
  5) Update the count parameters n_a^k, n_k^t as follows $n_a^k = n_a^k + 1$, $n_k^t = n_k^t + 1$.

// Run the chain
6) **Iterate** over a large number of iterations (e.g. 1000):
  7) **Iterate** over each word w in the corpus:
    8) Decrement the current word t's topic k and author a assignment counts as follows $n_a^k = n_a^k - 1$, $n_k^t = n_k^t - 1$.
    9) Sample a topic k and author a assignment for the word from
    $p(x = a, z = k|\mathbf{w}, \mathbf{z}_{-i}, \mathbf{x}_{-i}, A, \alpha, \beta) \propto \frac{n_k^t + \beta}{n_k + V\beta} \cdot \frac{n_a^k + \alpha}{n_a + K\alpha}$.
    10) Increment the new word/topic and topic/author counts as follows $n_a^k = n_a^k + 1$, $n_k^t = n_k^t + 1$.

// Compute model parameters
11) Estimate the model parameters as follows
$\hat{\phi}_k^t = \frac{n_k^t + \beta}{n_k + V\beta}$, and $\hat{\theta}_a^k = \frac{n_a^k + \alpha}{n_a + K\alpha}$, where $n_k = \sum_{t=1}^{V} n_k^t$, and $n_a = \sum_{k=1}^{K} n_a^k$.

Fig. 6. Gibbs sampling algorithm for ATM.

is assigned to topic k, and $n_a = \{n_a^k\}_{k=1}^K$ is the topic-author sum. In practice, parameter estimation is based on the procedure in Figure 6.

5.3 Perplexity

Perplexity is a common measure of the ability of a model to generalize to unseen data [Heinrich 2008]. It is defined as the reciprocal geometric mean of the likelihood of a test corpus given a model,

$$Perplexity = exp\left[- \frac{\sum_{m=1}^{M} \log p(w_m|\mathcal{M})}{\sum_{m=1}^{M} N_m} \right], \tag{10}$$

where N_m is the length of document m, $\mathcal{M}$ is the model, and w_m are the set of unseen words in document m. For all experiments described in Section 6, we used $\beta = 0.1$ and $\alpha = 50/K$.

In order to find the counts from a set of previously unseen documents, we:

(1) Divide the entire corpus into two groups, training and test sets. We randomly chose proportions of 90% training and 10% test documents.
(2) Run the inference algorithm on the training corpus.
(3) Run the inference algorithm on the test corpus, but "shift" the topic weights according to those obtained in Step 2 (training phase). More specifically, sample the topic/word and topic/document counts of the test corpus, but add the topic/word count of the training corpus to β before sampling.

5.4 Topic Models for Activity Modeling

To model human activities, we make an analogy between text documents and human location patterns. We replace words with location sequences, documents with days, topics with routines, and authors with users. The LDA model produces ϕ_t^k and θ_k^m, which represent the probability of location sequence t for each topic k, and the probability of topics k for each day m, respectively. Given these probability distributions, we can rank location sequences and days for each topic discovered, and determine routines which are discovered as topics.

The ATM model extends this interpretation to allow a varying set of routines to be discovered, this time with the emphasis of determining distributions of topics over authors, or routines followed by users. The ATM model produces ϕ_t^k and θ_k^a, which represent the probability of location sequence t for each topic k, and the probability of topics k for each user a, respectively. Given these probability distributions, we can again rank location sequences for each topic discovered. Furthermore, with this methodology we can also rank topics for users, resulting in the discovery of routines followed by users.

Based on our method, we set out to answer several questions:

—How can we use different types of topic models for location-driven human activity analysis, and more specifically, what type of topics do LDA and ATM discover?

—Are there specific activity patterns occurring on weekends versus weekdays?

—How do the topics discovered characterize the set of days and users in the dataset?

—Does the entropy of a user's location-routines have a meaning?

—How does the proposed method compare to clustering?

—Can the topic model methodology find changes in a user's daily location routines or discover meaningful groups?

We provide answers to these questions in the following section.

6. EXPERIMENTS AND RESULTS

In this section, we present our results motivated by the questions mentioned above. First we present the data used and describe the experiments used for model selection. We present the results of human location-driven activities from LDA and ATM. We then investigate daily patterns, compare our method to clustering, investigate users in terms of their location-entropy, and user our topic model method to discover groups of users' routines. Finally, the limitations of this work are covered.

6.1 Data

As summarized in previous sections, in the Reality Mining dataset [Eagle et al. 2009], the activities of 97 subjects were recorded by mobile phones over 491 consecutive days of data recording (January 1, 2004 to May 5, 2005). This comprises over 800 000 hours of data on human activity. The 97 subjects in the study are business and engineering students and staff of MIT living in a large

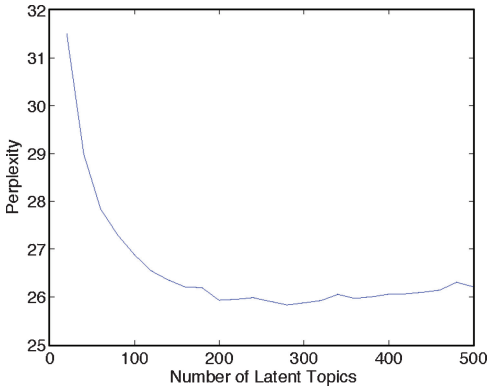


Fig. 7. Perplexity plot as a function of the number of latent topics, K. At $K = 200$, the perplexity mostly stabilizes to a low value.

geographical area covered by over 32000 cell towers. More precisely, 25 of the students in the dataset are labeled as Sloan business and the remaining 72 individuals are students and staff from the Media Lab. They work in offices with computers that have Bluetooth devices which can sense in a 5-10m radius. All privacy concerns of the individuals in the study have been addressed by the collectors of the data [Eagle et al. 2009]. For the experiments, we removed days which were entirely N (no reception), since they contained no useful information. The resulting dataset is still massive, amounting to 10 118 days, and over 242 800 hours of data. The set of days for experiments are visualized in Figure 1.

6.2 Model Selection for LDA

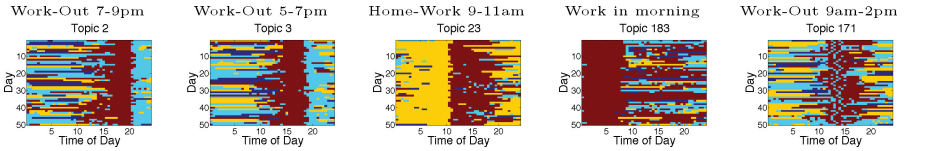
We use perplexity as a measure to determine the optimal number of latent topics, K. Due to space constraints, the detailed explanation of perplexity is given in the Appendix. We computed perplexity for LDA using K values from 20 to 500 with increments of 20. For all values of K, initialization was followed by 1000 iterations of the Gibbs sampling algorithm. The perplexity is plot over the number of latent topics in Figure 7. A drop in perplexity occurs at approximately $K = 200$ topics, after which the perplexity stabilizes. We choose $K = 200$ as the number of latent topics for the remaining experiments.

6.3 Routines Discovered with LDA

The LDA model successfully found latent topics over all users and days, and contain the dominating location routines. The unsupervised discovery of location-driven routines revealed different types of patterns, assigning intervals of days which follow characteristic trends to various topics with a probability measure. To illustrate the routines discovered, for each topic we rank the 4 most probable location sequences, ranked by $P(w|z)$, and show them in tables. For each topic, we also rank the 50 most probable days, ranked by $P(z|d)$, and visualize them in plots.

Table I. Varying Work Routines

Topic 2 - LDA		Topic 3 - LDA		Topic 23 - LDA		Topic 183 - LDA		Topic 171 - LDA	
Word	$p(w\|z)$	Word	$p(w\|z)$	Word	$p(w\|z)$	Word	$p(w\|z)$	Word	$p(w\|z)$
W W W 6	0.548	W W W 5	0.462	H H H 2	0.528	W W W 1	0.920	W W O 4	0.300
W O O 7	0.212	W W O 6	0.255	H W W 3	0.212	W W W 2	0.020	O W W 4	0.290
W W O 7	0.196	W O O 6	0.231	H H W 3	0.201	W W O 2	0.013	W O W 4	0.273
O O W 3	0.003	O W W 4	0.003	H H H 3	0.022	W O O 2	0.008	O O W 3	0.046



The table lists the 4 most probable location sequences ranked by $P(w|z)$ for topics 2, 3, 13, 23, 183, and 171. The visualizations beneath illustrate the corresponding 50 most probable days ranked by $P(z|d)$, entitled with the topic number and the semantic work routine.

In Table I, we illustrate the various types of work routines exhibited by listing the top location sequences with the corresponding topics' visualization of top days, ranked by $P(z|d)$.

Some interesting results are the following:

—Topics 2 and 3 in Table I capture "going from work to out in the evening" routines, at different time intervals. The most probable words for topic 2 are WWW6, which is being at work in timeslot 6 (5-7 pm) followed by going from work to out in timeslot 7 (7-9 pm) WOO7, WWO7. Topic 3 contains very similar top words, but in one timeslot sooner: it is characterized by being at work in timeslot 5 (2-5 pm) followed by going from work to out in timeslot 6 (5-7 pm). Beneath the table, we visualize the top days for those topics, and can see that the days in topic 3 contain a work to out transition at an earlier interval than in topic 2.

—Topic 23 captures a "going from home to work" routine between 9-11 am. The most probable words are "at home before 9 am", followed by HWW3, HHW3, which represent "going from home to work" transitions in timeslot 3 (9-11 am).

—Topic 183 captured "at work early in the morning", with the most probable words being WWW1 and WWW2 followed by transitions around 7-9 am.

—Topic 171 illustrates a "work to out fluctuation in the early afternoon" with top words containing work to out fluctuations in timeslots 3 and 4 (9 am-2 pm).

Note that in all these topics, the top few words account for over 90% of the probability mass, which suggests that the topics are discriminant of very characteristic patterns despite the inherent noise present in most days' data. This is possible due to the relatively large number of topics we use.

Other routines discovered are visualized in Figure 8 with their corresponding labels as the title. Note that these selected routines are just a few of the many meaningful topics discovered.

—Topic 15 captures a work to out to work routine which could correspond to a "lunch" break.

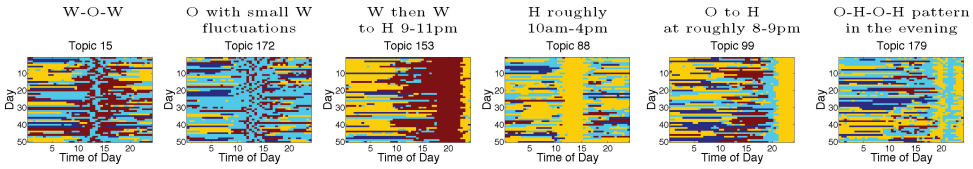
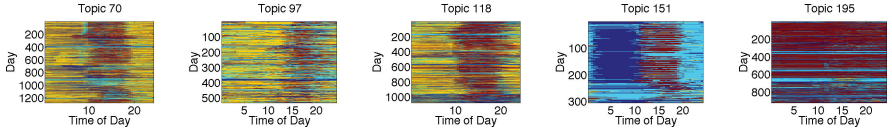


Fig. 8. A small subset of the routines discovered visualized for the top 50 days for each topic. The corresponding routine name is displayed above the discovered topics.

Table II. A Selection of Discovered ATM Routines

Topic 70 - ATM		Topic 97 - ATM		Topic 118 - ATM		Topic 151 - ATM		Topic 195 - ATM	
Word	$p(w\|z)$	Word	$p(w\|z)$	Word	$p(w\|z)$	Word	$p(w\|z)$	Word	$p(w\|z)$
H H H 1	0.346	H H H 1	0.289	H H H 1	0.221	N N N 1	0.193	W W W 1	0.332
H H H 8	0.151	H H H 4	0.159	W W W 5	0.16	W W W 5	0.175	W W W 5	0.121
H H H 7	0.128	H H H 3	0.14	W W W 6	0.14	W W W 4	0.125	W W W 4	0.12
H H H 6	0.064	H H H 2	0.113	W W W 4	0.113	N N N 2	0.116	W W W 2	0.099
User	$p(z\|a)$	**User**	$p(z\|a)$	**User**	$p(z\|a)$	**User**	$p(z\|a)$	**User**	$p(z\|a)$
95	0.286	62	0.348	54	0.234	14	0.320	26	0.639
11	0.226	57	0.209	29	0.219	43	0.100	27	0.618
15	0.213	63	0.163	10	0.160	78	0.089	58	0.578
39	0.205	75	0.156	85	0.153	8	0.081	24	0.526



Top location sequences are listed for selected topics, ranked by $p(w|z)$, as well as top users for these topics, ranked by $p(z|a)$. Beneath are plots for the top authors' days for given topics illustrating the routines discovered.

—Topic 172 captures out most of the day with very short work fluctuations occurring between 10 am-3 pm.

—Topic 153 captures working nonstop for at least 4 hours, then going home at approximately 10pm.

—Topic 88 captures home roughly 10am-3pm.

—Topic 99 captures out for a few hours at roughly 8pm, then arriving home at around 9 pm and staying home for the entire evening.

—Topic 179 captures an out-home-out-home routine, with each location occurring for a few hours in the evening.

6.4 Routines Discovered with ATM

For the ATM, we use the same model parameters as those for LDA. Specifically, $K = 200$, $\beta = 0.1$, $\alpha = 50/K$, and we run 1000 iterations of the Gibbs sampling algorithm. The results obtained by the ATM differ to those obtained by LDA. With the ATM, the routines capture users' routines, simultaneously taking into account users' identities and daily location routines. In contrast, the LDA model captures routines from the days in the dataset, disregarding users' identities.

In Table II, selected ATM results are listed. We include the top 4 location sequences for selected topics, ranked by the probability of a word given the topic,

$p(w|z)$. We also include the top 4 authors for the selected topics, ranked by the probability of the topic given a user, $p(z|a)$. Beneath the table, the plots entitled "Topic x", display all the days of users for which $p(z|a) > T_a$, where $T_a = 0.1$ ranked by users. We pick a selection of 5 from the 200 topics to demonstrate the routines obtained. Note that each user has a different number of recorded days in the dataset, and each topic has differing number of users with $p(z|a) > T_a$, which explains the varying number of days plot for each topic.

—In Topic 70, the top words are "being at home in the early morning (HHH1) and evening from 5 pm onwards" (HHH6, HHH7, HHH8). Users whose daily lives most often evolve around this routine are users 95, 11, 15 and 39 who characterize this topic with similar probabilities.

—Topic 97 discovered a "being at home early in the day" (HHH in timeslots 1-4). Users 62, 57, 63 and 75 display this routine most frequently, though not everyday, as seen by the lower $p(z|a)$ for users 63 and 75. In the corresponding plot, we can see a general "being at home in the mornings and afternoons" routine, though not everyday.

—Topic 118 found a "being at home" in the morning routine (HHH1) co-occurring with "being at work 11am-7pm" (WWW4, WWW5, WWW6). Users 54, 29, 10, and 85 most frequently follow this daily life pattern.

—Topic 151 captures "no reception in the morning" (NNN1 and NNN2) co-occurring with "being at work in the afternoon" (WWW4, WWW5). Users 14 and 43 most strongly follow this routine.

—Topic 195 discovered a "being at work throughout the day" routine (WWW in timeslots 1,2,4,5), which is very frequently followed by users 26, 27, 58, and 24, seen by their high $p(z|a)$ and their daily lives visualized in the figure. This could potentially be the discovery of "users who live on campus".

Comparing the routines obtained with LDA and ATM, we note that the ATM produces topics with top words that do not account for the probability mass as strongly as they do in LDA. Also, note that none of the top users shown for the topics are the same. This suggests that the ATM is preferring certain users versus others for these topics. Also, note that for some topics, authors are very characteristic (high $p(z|a)$), while for other topics this is not the case. Overall, the ATM learned topics that are more general than the ones with LDA, with the advantage of learning the author-topic distributions. We lose discrimination in the topics (routines discovered) with ATM but this is traded off for learning author distributions.

6.5 Daily Patterns

Our methods allow to extract daily patterns that are meaningful according to the day type and seen as a mixture. We now discuss these two aspects.

6.5.1 *Weekend and Weekday-Like Routines.* On a weekly level, some trends characteristic of weekends versus weekdays appeared with the routines discovered by LDA. For example, topics 182 and 122, plot in Figure 9,

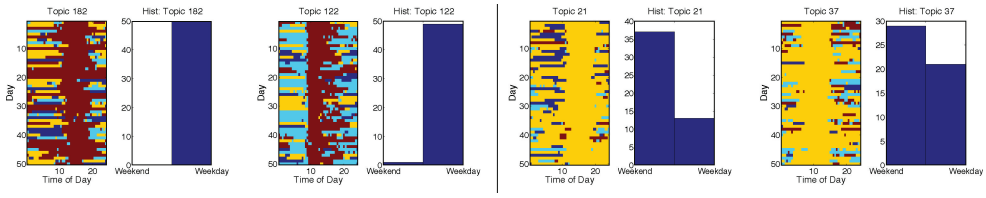


Fig. 9. Weekend-dominant versus weekday-dominant routines discovered by LDA. The visualizations entitled "Topic x" show the top 50 days, ranked by $P(z|d)$, for topic x. The plots entitled "Hist:Topic x" are counts of whether the most probable days in topic x correspond to Weekends or Weekdays. It can be seen that the top 50 days for topics 182 and 122 almost entirely correspond to weekdays. The majority of the most probable days of topics 21 and 37 correspond to weekends.

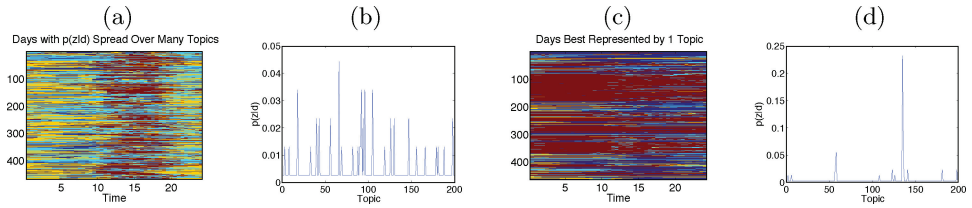


Fig. 10. (a) Days which are described by many topics in LDA. (b) $p(z|d)$ plot for a given day which is described with low probability by many topics. (c) Days which are well described by a single topic. (d) $p(z|d)$ plot for a given day which is well described by a single topic.

demonstrate routines which dominated on weekdays and topics 21 and 37 demonstrate routines which tend to dominate on weekends. Each visualization of the most probable days per topic, entitled "Topic x" is followed by a histogram, entitled "Hist: Topic x", which counts whether the topic's 50 top days correspond to weekends or weekdays. We can see a "being at work" routine in topics 182 and 122 corresponds to weekday trends, and "being at home" during the day corresponds mostly to weekend behavior, though some weekdays also demonstrate this routine, perhaps corresponding to holidays or days off.

6.5.2 *Days as Mixture of Topics.* One fundamental question that arises is: how evident is the "mixture of topic" assumption in our data. Are days about one topic or more? Our LDA methodology also allows us to find days which vary over many topics, and days that are best represented by a few topics. On one hand, by looking at days for which $p(z_i|d) <= T_L, \forall i = 1 : K$ where T_L is a small value (set here to 0.05), we can find days that are not highly probable for any topic and thus are distributed with low probabilities over many topics. These days are visualized in Figure 10(a), and the probability distribution of topics given a particular day which is not highly probable for any topic is shown in Figure 10(b). On the other hand, we also find days that are best represented by a few topics by collecting days for which $p(z_i|d) > T_H$, for a given i, where $T_H = 0.15$. These days are visualized in Figure 10(c) and illustrate days which are best characterized by few topics. In Figure 10(d), we plot the probability distribution of topics given a day which is well represented by a single topic. The thresholds T_L and T_H were picked in order to depict data on the order of 500 days. Comparing Figures 10(a) and (c) we can differentiate between days

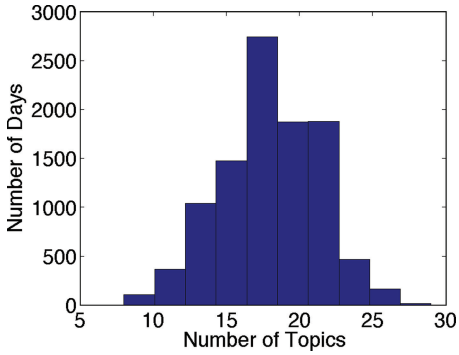


Fig. 11. Histogram of number of 'dominating' topics per day for the LDA model.

following a rich set of routines and days lacking in variety in terms of location patterns. Those with highly varying routines, generally require more topics to capture their structure.

In Figure 11, we show a histogram of the number of "dominating" topics per day. We compute the number of topics composing at least 50% of the probability mass of each day in the study, and plot a histogram of the results. In general, all days are well described by fewer than 30 topics. Thus, at most 15% (30/200) of the topics can describe the probability mass of any day in the dataset. On the lower end of the histogram, very few days are described by less than 10 topics (35 days, or 0.3% of the days in the dataset). The same can be observed for high number of topics, very few days require 25 or more topics to be well defined (180 days, or 1.8% of the days in the dataset). The average number of topics in the study is 18 topics. Therefore, even though people typically follow very routine daily lifestyles, as found in Gonzalez et al. [2008], their daily location routines are true mixtures, involving a mixture of around 20 topics on average to define over 50% of the probability mass of the day.

6.6 Clusters versus Topics

One basic question is whether the topic model discovers groups of days different than classical clustering algorithms would. To investigate this, we compare the results of routine discovery from the k-means clustering algorithm [Duda et al. 2000] to those obtained by LDA. For this task we run k-means with 50 clusters and compare the results to LDA with $K = 50$ topics. The input to K-means is the fine-grain location representation (in binary). Both algorithms are initialized randomly. The results are illustrated in Figure 12. Results are presented for a small number of topics for simplicity of visualization and analysis. We observe that k-means finds very general routines occurring 'broadly' over the entire day. In contrast, LDA finds topics with patterns occurring over parts of the day, as well as days with specific transition patterns occurring at a given time, such as those shown in Figure 12(c) and (d). Furthermore, LDA discovers several routines such as the one visualized in Figure 12(e), where alternating locations occur for varying time durations, which are not found by k-means. They are discovered with LDA due to the exchangeability of words assumption [Blei et al.

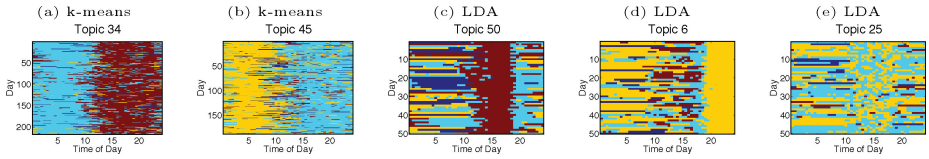


Fig. 12. K-means clustering versus LDA topic discovery. (a) and (b) illustrate two typical clusters obtained by k-means. (c)-(e) illustrate three topics obtained by LDA. K-means clustering discovers very general routine discovery, occurring over the entire day. Topic discovery results in the probabilistic ranking of discriminative words. These discriminative words result in the determination of routines occurring dominantly over parts of the day. Further, transitional patterns, such as the home-out fluctuations in (e), can be found with LDA, but not with K-means.

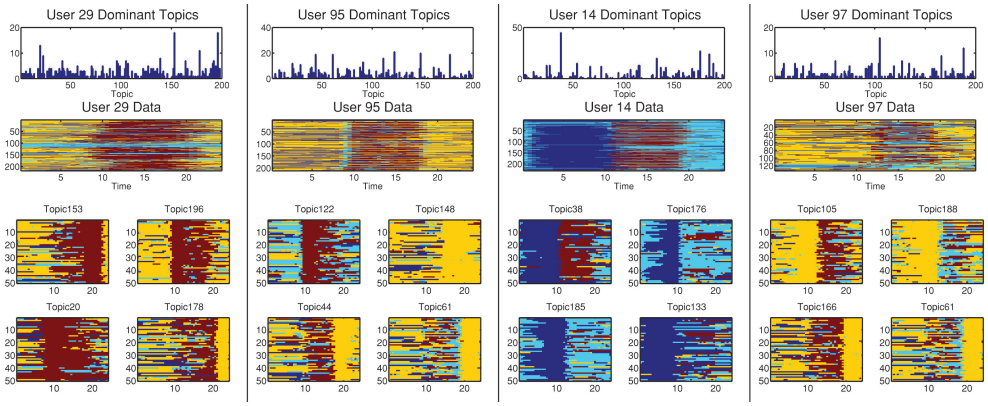


Fig. 13. Individual user analysis. The histograms "User x Dominant Topics" demonstrate dominant topics for users x. Plots "User x Data" corresponds to the raw input location data of user x. The four topics are the four dominating routines for user x.

2003], which cannot be found using basic clustering techniques which take the exact occurrence of labels (here location) into account for comparison between data vectors (here days). More generally, the advantage of topic models over traditional clustering methods are: (1) soft clustering of days, and (2) meaningful word distributions as the representation of topics. Concretely in our work, the LDA model results in (1) probabilistic distributions of days given all topics whereas k-means assigns only one cluster per day, and (2) discriminative location sequences per topic characterizing human routines. This information is very useful as we know the precise location transitions which characterize the human routine as well as the timestamp, giving the routine's time details.

6.7 Modeling Individual Users with Topics

6.7.1 *LDA-Based User Analysis.* We can also examine the topic distributions over users with LDA. For each user x's day d_x, we count the topics for which the ranked probability of the topics given the day, $p(z|d_x)$ is greater than T_L, aggregate for all the user's days and illustrate them in the histogram entitled "User x Dominant Topics" in Figure 13. Some users' days are expressed well by a few topics, other users have a rich set of varying routines that are

expressed as a mixture of many topics. For example, noting the varying y-axis scales, user 14 has a very high probability of a few topics for most days, whereas users 29, 95, and 97's days are expressed as a mixture over many topics. We plot the users' location data in the plots entitled "User x Data". Each user has a different number of days ($y-$axis), since they have varying number of days after removing fully *no reception* days. Beneath the users' days are the four topics which dominate the given users' daily activities. For instance, the four topics dominating user 29's daily routines are topics 153, 196, 20, and 178. User 29's dominating routines are "being at work" routines, as well as "being at work late in the evening". Looking at "User 29 Data", we can confirm that user 29 does work a lot, especially late in the evening. User 29's daily activities are thus a mixture over several topics, as can be seen by the histogram "User 29 Dominant Topics". User 95's most common routines are "arriving to work before 10 am from an O location", and "being at home in the afternoons and evenings". Looking at user 95's location data, we can see this user is at home in the morning then goes to an O location, perhaps for breakfast or the gym, then goes to work and home in the evenings. User 14 mostly has no reception in the mornings, followed by being at work during the day, as seen by the dominant topic 38 dominating most of his/her daily activities. The user is mostly out in the evenings. It appears that this user's home label is missing, and she/he either turns the phone off while sleeping, or loses reception early in the morning. User 14's dominant topics are less of a mixture over several topics than users 29, 97 and 95. User 97's routines are predominantly "at home in the morning and evenings".

6.7.2 *ATM-Based Analysis.* We can also analyze individual user's daily structure with the topics discovered by ATM. The plots in Figure 14 illustrate how well users' daily routines are described by the topics discovered by the ATM. In Figure 14(a), we plot the number of dominating topics composing 70% of the probability mass for each user. Most users' daily routines are described well by under 17 topics. Users with no data are not considered. Some users are well characterized by a few topics, others require more. In Figure 14(b), we plot 2 individual users that vary in their daily routine distributions over topics. User 14 (also discussed for the LDA case) is well characterized by 2 topics, whereas user 65 is characterized by a mixture of many topics. In Figure 14(c) and (d), we plot the days of 4 users whose daily life patterns are described well by a mixture of 2 topics, and 4 users whose daily routines are described by a mixture over many topics. Visible differences between these users' lifestyles are that users 14, 26, 59, and 60 follow very regular, non-varying lifestyles, which are captured well by a few topics, whereas users 83, 89, 9, and 65 have varying daily routines. We also plot a histogram of the number of users whose lives are characterized by various number of topics in Figure 14(e). According to the ATM analysis, many users can actually be well characterized by few topics. 10 users can be well characterized by fewer than 4 topics, and 18 users by fewer than 5 topics, demonstrating that a significant portion of users have very repetitive non-varying lives. Fewer users have more highly varying lifestyles, as seen by the higher end of the histogram. 11 users can be well described by more than 12 topics. In Figure 14(f), we plot the entropy for each user, computed

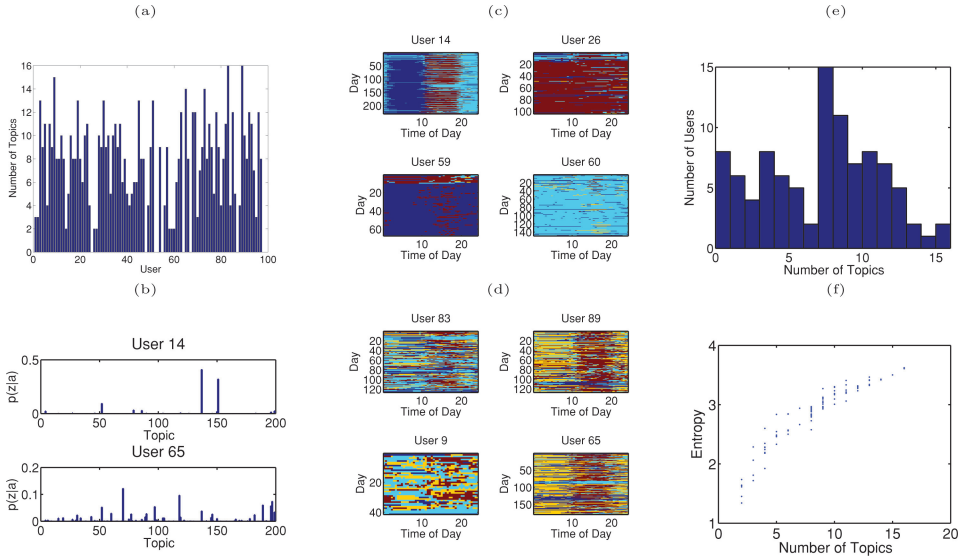


Fig. 14. ATM Results. (a) Plot of the number of dominating topics for each user. Most users' daily routines are described well by less than 20 topics. (b) Topic distribution for users 14 and 65. User 14 is well characterized by 2 topics, whereas user 65 is characterized by a mixture of topics. (c) The days of 4 users whose daily life patterns are described well by a mixture of 2 topics. (d) The days of 4 users who are characterized by a mixture of many topics. (e) A histogram of the number of users as a function of the number of topics. (f) Number of topics plot as a function of entropy for each user, showing an approximate linear relationship between the two measures.

on the topic distribution, as a function of the number of dominating topics. Each data point represents an individual user. We can see that the number of topics as a function of entropy is about linear, suggesting that number of dominating topics is indeed a good measure of user entropy and variation in daily activities.

6.8 Finding Variations in Individual's Lives Over Time

In order to find variations in a user's daily routines over time, we compute the Bhattacharya distance between consecutive days of a user, $BD = \sqrt{1 - \sum_{z \in K} \sqrt{p(z)q(z)}}$, where $p(z) = p(z_k|d_i)$ and $q(z) = p(z_k|d_{i+1})$, where d_i and d_{i+1} indicate consecutive days. The Bhattacharya distance measures the similarity of two discrete probability distributions. The more similar two probability distributions are, the closer the sum of products term in BD will be to 1. The smaller the overall BD term, the more similar two probability distributions, with a minimum value of 0. If two probability distributions differ significantly, the sum of product terms will be smaller, and the resulting BD expression will approach its maximum value of 1.

This measure proves to be useful in finding changes in a user's routines over time, as shown for users 2 and 24 in Figure 15. The bottom figures are the Bhattacharya distance plot as a function of day. The top figures are the set of days for both users, corresponding to the days in the Bhattacharya distance plots above. Note that here days are on the x-axis and time of day on the y-axis.

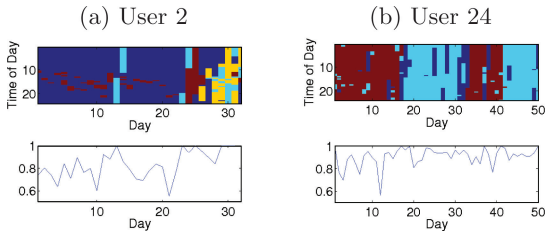


Fig. 15. The difference between routines over consecutive days plot for (a) User 2 and (b) User 24. The top plots visualize several days with the time of day in hours ($y-$axis) vs. consecutive days ($x-$axis). The bottom plots correspond to Bhattacharya distance between pairs of consecutive days.

Figure 15(a) shows that at day 13-14, there is a change in the user's activities where they go from "N in the morning and evening" to "being out". This change is captured in the Bhattacharya distance, where the first peak to 1 occurs. The second peak occurs at day 23, where there is again a large variation in the user's routines, this time consisting of "W" routines for a few days followed by "H" routines for some days. User 24's changes in daily routines are also captured by the Bhattacharya distance, where the measure peaks to 1 at day 17, 19, 37, and 42.

We noted that this measure is sometimes sensitive to days for which large variations don't occur. This is due to several topics accounting for similar routines, likely to happen given the relatively large number of topics. For example, there are more than one topic displaying "being at work in the morning" or "being out during the day". Given the stochastic nature of Gibbs' sampling, 2 very similar days could have differing topic distributions. Thus, there are certain days which are very similar in structure but have larger than expected Bhattacharya distance measures. This could be a result of the potential inadequacy of the perplexity measure in model selection, which is discussed further in Section 6.10.

6.9 Group-Level Routines

6.9.1 *LDA-Based Analysis.* Using the LDA methodology, we can determine users that exhibit certain patterns. We find topics that display certain routines we wish to inspect such as "working late". There may be more than one topic displaying a similar routine. We use LDA to rank top days for these routines, and then count the number of times these days are highly ranked for each user. The individuals displaying these routines the most are listed as Users in Table III.

As three examples, we can first find users that "go to work early". There are two topics that display this routine, one of them displaying "going to work from home" (Topic 196) and the other one "going to work from other" (Topic 122). In both cases the user is at work by timeslot 3 (9 am-11 am). Users 29, 4, 69, and 10 arrive to work early from home, as seen in Topic 196 in Table III. Users 95, 41, and 13 arrive to work early from O. Second, we can also find users that work very late before going home. Topic 153 dominates for users 29, 10, and 78 who often stay at work late (5-9 pm) and arrive home after 9 pm. Finally,

Table III. Users that Follow the Specific Routines of "Going to Work Early", "Working Late then Going Home", and "Turning off Their Phones (or no Reception) in the Evening

Topic 196 - LDA W early from H			Topic 122 - LDA W early from O			Topic 153 - LDA Work late then H			Topic 104 - LDA Phone off or N in evening		
User	Word	$p(w\|z)$	User	Word	$p(w\|z)$	User	Word	$p(w\|z)$	User	Word	$p(w\|z)$
29	WWW3	0.61	95	WWW3	0.59	29	WWW7	0.44	90	NNN8	0.65
4	HHW2	0.17	41	WWW4	0.15	10	WWW6	0.32	94	NNN7	0.09
69	HWW2	0.12	13	OWW2	0.11	78	WHH8	0.10	42	ONN7	0.05
10	HWW3	0.04		OOW2	0.09		WWH8	0.10	37	OON7	0.04
	WWW2	0.01		HOW2	0.02		WHO8	0.01		HNN7	0.04

We list users that follow these routines more than usual.

Table IV. Routines Discovered by ATM Showing Business Student Activities

Topic 104 - ATM		Topic 139 - ATM		Topic 145 - ATM		Topic 146 - ATM	
User & Type	$P(z\|a)$	User & Type	$P(z\|a)$	User & Type	$p(z\|a)$	User & Type	$p(z\|a)$
38 Sloan	0.040	12 Sloan	0.022	1 No label	0.187	4 Sloan	0.130
50 Sloan	0.036	82 Sloan	0.011	84 Sloan	0.125	73 Sloan	0.100
96 student	0.033	66 Sloan	0.011	50 Sloan	0.111	36 Sloan	0.091
82 Sloan	0.027	69 newgrad	0.009	42 Sloan	0.088	79 Sloan	0.085
Word	$p(w\|z)$	Word	$p(w\|z)$	Word	$p(w\|z)$	Word	$p(w\|z)$
H H H 8	0.115	H H W 7	0.164	O O O 5	0.258	H H H 5	0.158
H H O 8	0.099	H W W 7	0.135	O O O 4	0.241	H H H 1	0.130
H H H 7	0.097	W H H 6	0.119	O O O 3	0.149	H H H 4	0.127
H O O 8	0.079	O W W 3	0.052	O O O 6	0.084	N N N 4	0.100

Displayed topics are discovered to be dominating for business (Sloan) students.

we discover users that possibly turn off their phones (have no reception) in the evenings. Users 90, 94, 42, and 37 often have no reception after 7 pm.

6.9.2 *ATM-Based Analysis.* While groups of users that share a routine can be discovered by LDA, ATM is in general better suited for this task. By finding several users that display given routines, we can identify small groups of individuals displaying similar behavior. With the ATM methodology, certain topics exhibit routines that dominate for mostly business students. In Table IV, we display 4 of the routines that were discovered for business students. We list the top users along with their student types. We also display the top location sequences for these topics.

—Topic 104 dominates for users 38, 50, 96, and 82, three of which are Sloan students. These students tend to go out from home late in the evenings in timeslot 8 (9-12 pm).

—Topic 139 is dominant for users 12, 82, 66, and 69, three of which are Sloan students. These users go from home to work in timeslot 7 (7-9 pm) or from work to home in timeslot 6 (5-7 pm).

—Topic 145 occurs mostly for users 1, 84, 50, and 42, three of which are Sloan students. These users are often at O locations throughout the day.

—Topic 146 occurs dominantly for four business students. They are often at home in timeslots 4 and 5 (11 am-5 pm).

6.10 Limitations

While we have shown that many insights about routines can be obtained with our approach, our work has a number of limitations. The first one is the scope of users for which data was collected. The users in this study are all MIT students and staff, and their daily routines are clearly not representative of the whole society. However, their daily routines are expected to be similar, for the most part, to other university students and staff as well as working professionals. Further, the fact that two types of students and staff were used as the population for which data is collected (engineering vs. business) makes the dataset more representative.

Another limitation is the way we select the number of topics. For LDA, perplexity is used as a measure to determine performance on unseen data. However, perplexity is not a "perfect" measure for model selection, since similar resulting topics is not considered in the perplexity computation. In practice, smaller values of K would have resulted in less "duplication" of topics (as we observed in our preliminary work [Farrahi and Gatica-Perez 2008b]) but also the topics become more general. Overall, perplexity is not perfect for model selection, though other ways of choosing model parameters are not much better, and the issue of model selection for topic models is an active problem [Blei et al. 2004]. For ATM, this parameter was chosen to be the same as that for LDA for comparison purposes, but there is obviously no guarantee that such value would have been chosen from perplexity experiments.

7. CONCLUSION

We have presented a novel framework for location-driven routine discovery using probabilistic topic models. Using a massive dataset collected by 97 users' mobile phones over a period of 16 months from the Reality Mining project, we successfully discover routines characteristic of days and individuals in the study in an unsupervised manner. We have proposed a method to represent location sequences, and incorporated this into the LDA and ATM topic models. The resulting distributions of words for latent topics, as well as topics given days, and topics given users, reveal hidden structure of routines which we use to perform varying tasks, including finding users or groups of users that display given routines, and determining times when certain events or changes in events occur.

We have several ideas for extending this work. We would like to investigate the proximity dataset, obtained by Bluetooth information of the users' mobile phones. By considering who people are in contact with at various times of the day, we could perform various tasks including differentiating groups of co-workers, officemates, or friends. As large-scale mobile sensing becomes more feasible, we would like to test our method on other datasets [Kiukkonen et al. 2010]. Recent work has investigated parallel versions of LDA, which make the model scalable to handle large numbers of users and days [Wang et al. 2009]. We are also interested in investigating other topic models, such as hierarchical topic models [Blei et al. 2004], with the goal of inferring the number of latent topics within the model. Furthermore, we are interested to look at models

which could account for varying routine time intervals, specifically analyzing routines on varying timescales, such as hourly, daily and weekly.

ACKNOWLEDGMENTS

We thank Nathan Eagle (Santa Fe) and Alex Pentland (MIT) for sharing the data, and Sileye Ba (IDIAP) for his insightful comments.

REFERENCES

AKOUSH, S. AND SAMEH, A. 2007. Mobile user movement prediction using bayesian learning for neural networks. In *Proceedings of the ACM International Wireless Communications and Mobile Computing Conference (IWCMC)*. 191–196.

BLEI, D. M., GRIFFITHS, T. L., JORDAN, M. I., AND TENENBAUM, J. B. 2004. Hierarchical topic models and the nested Chinese restaurant process. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

CHOUDHURY, T. AND PENTLAND, A. 2003. Sensing and modeling human networks using the sociometer. In *Proceedings of the IEEE International Symposium on Wearable Computers (ISWC)*. 216.

CHOUDHURY, T., PHILIPOSE, M., WYATT, D., AND LESTER, J. 2006. Towards activity databases: Using sensors and statistical models to summarize people's lives. *IEEE Data Engin. Bull*. 49–58.

DUDA, R., HART, P., AND STORK, D. 2000. *Pattern Classification* 2nd Ed. Wiley-Interscience.

EAGLE, N. AND PENTLAND, A. 2009. Eigenbehaviors: Identifying structure in routine. *Behav. Ecol. Sociobiol. 63,* 7, 1057–1066.

EAGLE, N., PENTLAND, A., AND LAZER, D. 2009. Inferring social network structure using mobile phone data. *Proc. Nat. Acad. Sci. 106,* 36, 15274–15278.

FARRAHI, K. AND GATICA-PEREZ, D. 2008a. Discovering human routines from cell phone data with topic models. In *Proceedings of the IEEE International Symposium on Wearable Computers (ISWC)*.

FARRAHI, K., AND GATICA-PEREZ, D. 2008b. What did you do today? Discovering daily routines from large-scale mobile data. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

GONZALEZ, M. C., HIDALGO, C. A., AND BARABASI, A.-L. 2008. Understanding individual human mobility patterns. *Nature 453,* 7196, 779–782.

GRIFFITHS, T. L. AND STEYVERS, M. 2004. Finding scientific topics. *Proc. Nat. Acad. Sci. USA 101* Suppl 1, 5228–5235.

HEINRICH, G. 2008. Parameter estimation for text analysis,. Tech. rep., University of Leipzig.

HIGHTOWER, J., CONSOLVO, S., LAMARCA, A., SMITH, I., AND HUGHES, J. 2005. Learning and recognizing the places we go. In *Proceedings of the Workshop on Ubiquitous Computing (UbiComp)*. 159–176.

HUYNH, T., FRITZ, M., AND SCHIELE, B. 2008. Discovery of activity patterns using topic models. In *Proceedings of the Workshop on Ubiquitous Computing (UbiComp)*. 10–19.

INTILLE, S. S., LARSON, K., TAPIA, E. M., BEAUDIN, J. S., KAUSHIK, P., NAWYN, J., AND ROCKINSON, R. 2006. Using a live-in laboratory for ubiquitous computing research. In *Proceedings of the Workshop on Pervasive Computing*. 349–365.

KIUKKONEN, N., BLOM, J., DOUSSE, O., GATICA-PEREZ, D., AND LAURILA, J. 2010. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proceedings of the ACM International Conference on Pervasive Services (ICPS)*.

KRUMM, J. AND HORVITZ, E. 2006. Predestination: Inferring destinations from partial trajectories. In *Proceedings of the Workshop on Ubiquitous Computing (UbiComp)*.

LARSON, K. AND INTILLE, S. Placelab website. http://architecture.mit.edu/house_n/data/placelab/placelab.htm.

LETCHNER, J., FOX, D., AND LAMARCA, A. 2005. Large-scale localization from wireless signal strength. In *Proceedings of the Natural Conference on Artificial Intelligence (AAAI)*. 15–20.

Liao, L., Fox, D., and Kautz, H. 2006. Location-based activity recognition. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*. 787–794.

Ling, R. and Donner, J. 2009. *Mobile Communication: Digital Media and Society Series*. Polity Press.

Logan, B., Healey, J., Philipose, M., Tapia, E. M., and Intille, S. S. 2007. A long-term evaluation of sensing modalities for activity recognition. In *Proceedings of the Workshop on Ubiquitous Computing (UbiComp)*. Lecture Notes in Computer Science, vol. 4717, Springer, 483–500.

MIT Technology Review. 10 emerging technologies 2008. http://www.technologyreview.com/specialreports/specialreport.aspx?id=25.

Monay, F. and Gatica-Perez, D. 2007. Modeling semantic aspects for cross-media image retrieval. *IEEE Trans. Patt. Anal. Mach. Intell.*

Niebles, J., Wang, H., and Fei-Fei, L. 2006. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vision*.

Otsason, V., Varshavsky, A., Lamarca, A., and de Lara, E. 2005. Accurate gsm indoor localization. In *Proceedings of the Workshop on Ubiquitous Computing (UbiComp)*. Springer, Berlin, 141–158.

Patterson, D., Liao, L., Fox, D., and Kautz, H. 2003. Inferring high-level behavior from low-level sensors. In *Proceedings of the Workshop on Ubiquitous Computing (UbiComp)*. 73–89.

Pritchard, J., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics. 155*, 945–959.

Quelhas, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., and Tuytelaars, T. 2007. A thousand words in a scene. *IEEE Trans. Patt. Anal. Mach. Intell.*, 1575–89.

Reddy, S., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. 2008. Using mobile phones to determine transportation mode. In *Proceedings of the International Symposium on Wearable Computers (ISWC)*.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. 2004. The author-topic model for authors and documents. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. 487–494.

Sohn, T., Varshavsky, A., Lamarca, A., Chen, M., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W., and de Lara, E. 2006. Mobility detection using everyday gsm traces. In *Proceedings of the Workshop on Ubiquitous Computing (UbiComp)*. 212–224.

Tapia, E., Intille, S., Haskell, W., Larson, K., Wright, J., King, A., and Friedman, R. 2007. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In *Proceedings of the International Symposium on Wearable Computers (ISWC)*. 37–40.

Tapia, E. M., Intille, S. S., and Larson, K. 2004. Activity recognition in the home setting using simple and ubiquitous sensors. In *Proceedings of PERVASIVE*. 158–175.

Wang, Y., Bai, H., Stanton, M., Chen, W.-Y., and Chang, E. Y. 2009. PLDA: Parallel latent dirichlet allocation for large-scale applications. In *Proceedings of AAIM*. 301–314.

Wren, C., Ivanov, Y., Kaur, I., Leigh, D., Westhues, J., Wren, C. R., Ivanov, Y. A., Kaur, I., Leigh, D., and Westhues, J. 2007. Socialmotion: Measuring the hidden social life of a building. In *Proceedings of the International Symposium on Location- and Context-Awareness (LoCA'07)*. 85–102.

Received March 2010; accepted May 2010