Probabilistic Mining of Socio-Geographic Routines From Mobile Phone Data

Katayoun Farrahi, *Member, IEEE*, and Daniel Gatica-Perez, *Member, IEEE*

***Abstract*—There is relatively little work on the investigation of large-scale human data in terms of multimodality for human activity discovery. In this paper, we suggest that human interaction data, or human proximity, obtained by mobile phone Bluetooth sensor data, can be integrated with human location data, obtained by mobile cell tower connections, to mine meaningful details about human activities from large and noisy datasets. We propose a model, called bag of multimodal behavior, that integrates the modeling of variations of location over multiple time-scales, and the modeling of interaction types from proximity. Our representation is simple yet robust to characterize real-life human behavior sensed from mobile phones, which are devices capable of capturing large-scale data known to be noisy and incomplete. We use an unsupervised approach, based on probabilistic topic models, to discover latent human activities in terms of the joint interaction and location behaviors of 97 individuals over the course of approximately a 10-month period using data from MIT's Reality Mining project. Some of the human activities discovered with our multimodal data representation include "going out from 7 pm–midnight alone" and "working from 11 am–5 pm with 3–5 other people," further finding that this activity dominantly occurs on specific days of the week. Our methodology also finds dominant work patterns occurring on other days of the week. We further demonstrate the feasibility of the topic modeling framework for human routine discovery by predicting missing multimodal phone data at specific times of the day.**

***Index Terms*—human activity, human mobility, Reality Mining, topic models.**

I. Introduction

CELL phones are rapidly emerging as the ultimate multimodal sensor of human dynamics and behaviors [11]. Equipped with GPS, Bluetooth, accelerometers, cameras, and microphones, current phones have the potential of tracing multiple forms of data at scales previously unattainable. This data has the potential of enabling the design of new human-centered applications related to people's daily life, thus opening a whole scope of problems in multimodal integration and ubiquitous computing [4], [16], [19], as well as enabling the understanding of human interactions, movements, and behaviors, and how these impact each other, as never before.

Manuscript received November 29, 2009; revised January 27, 2010; accepted April 11, 2010. Date of publication May 3, 2010; date of current version July 16, 2010. This work was supported by the Swiss National Science Foundation through the MULTI project. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Vikram Krishnamurthy.

The authors are with the Idiap Research Institute, CH-1920 Martigny, Switzerland, and also with the Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland (e-mail: kfarrahi@idiap.ch; gatica@idiap.ch.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2010.2049513

Two fundamental problems in this domain relate to routine modeling: how to *discover* recurrent patterns in a person's life from multimodal data like proximity, location, and motion, and how to *predict*, based on the knowledge of a person's routines, her most likely routines at any given time. On one hand, pattern discovery via unsupervised learning is often a necessity, given the potentially large number of relevant routine patterns of an entire population and the huge amount of unlabeled data that can be recorded with a phone over time [7], [8]. On the other hand, predictions from aggregated user observations are, arguably, some of the most useful outcomes of routine modeling, by inferring both where and with whom a user would most likely be in the future (for anticipation) or would most likely have been in the past (for cases of missing data).

While recent works have started to analyze both problems from location or proximity data—discovery and prediction in [7], discovery in [8]—one aspect that has not been investigated in depth is the role of multimodal integration in large-scale routine analysis. More specifically, how does the joint use of multiple modalities (e.g., location and proximity to others) enhance the understanding of a person's routines, and how can this be efficiently represented and automatically inferred? Proximity to known people (as a coarse approximation of face-to-face interaction) adds a rich element of social context that is very useful to complement or disambiguate many situations in daily life. For instance, being at home alone or with a large group having a party represent entirely different social situations, that would be nevertheless identical from the sole perspective of location. Such finer descriptions of routines based on multiple cues are clearly important to characterize users and their habits.

This paper presents an approach for large-scale unsupervised learning and prediction of people routines through the joint modeling of human location and proximity interactions. Our work has four contributions:

1) We present an approach to jointly model a user's location, interactions, and time data in a manner suitable for robust human activity mining from large-scale noisy data. We propose a bag of multimodal behavior, that integrates the modeling of variations of semantic location over multiple time-scales, and the modeling of interactions types from Bluetooth proximity. Our representation is simple yet robust to noisy and incomplete real-life mobile data.
2) We present an analysis of the proximity interactions occurring in the Reality Mining data [7] which depicts MIT Media Lab and business students, considering both durations of interactions with fellow lab-mates as well as all other Bluetooth devices.
3) We use a probabilistic topic model, namely Latent Dirichlet Allocation (LDA), to mine the dominant multi-

modal human activities occurring in the Reality Mining data, including typical human activities such as "being at home in the morning with another person." Upon closer analysis of the results, we are able to find routines occurring dominantly on certain days of the week to inform us of activities such as "being out Friday evening with a large group of lab mates."

4) We present a method to predict missing multi-modal sensor data, in this case joint location-proximity data over several hour intervals. The prediction task further confirms the feasibility of the joint location-proximity routines discovered as topics for data prediction.

This paper is organized as follows. In Section II, we present the most related works on large-scale human sensor data with focus of human activity modeling. We then present our framework in detail in Section III, followed by our experimental results and analysis in Section IV. Finally, the paper is concluded with some ideas for future work.

II. Related Work

The mobile phone is a very unique device continuously capturing our location, interaction, communication, and motion traces continuously left behind in our daily lives [16]. Researchers are just beginning to understand the implications of such data collections for fields ranging from epidemiology [24] to dynamical network analysis [14]. The research most relevant to ours is in the field of human activity modeling.

There is an increasing body of work on activity recognition using various types of wearable sensors (not involving mobile phones). For example, in [21], wearable electronic badges measure the amount of face-to-face interaction, conversational time, physical proximity to other people, and physical activity levels in order to capture individual and collective patterns of behavior. Their goal is to understand how patterns of behavior shape individuals and organizations. Other authors [15] use two wearable sensors, one placed on the right hip and one on the right wrist of a person, to recognize user daily routines such as "driving a car," and "washing hands." The method uses topics models and is tested on a few weeks of data obtained by one user. Their technique, however, would not be directly applicable to mobile sensor data since it uses body part sensitive human motion features of the wrist and hip as opposed to features directly obtainable by mobile phones that can be worn in pockets, bags, or backpacks. Recently, mobile phones have been modified to capture nonlinguistic speech attributes [19], [20]. These nonverbal speech features have been used for sound classification (for example music versus voice) and for the discovery of sound events [19]. In [20], these features, in addition to others obtained by mobile phone sensors, are used to characterize social interactions such as personal relationships at the workplace or in private.

Mobile phone call data has been analyzed at large-scales in [3] and [11] to understand human dynamics. Human mobility patterns have been modeled from location data obtained whenever phone calls were made in [11], to find that human trajectories are highly regular in terms of both temporal and spatial characteristics. In [3], phone call data has been used to study the mean collective behavior of humans at large scales, focusing on the occurrence of anomalous events. The authors also investigate patterns of calling activity at the individual level and model the individual calling patterns (time between phone calls) as heavy tailed. In [5], missing data in activity logs are filled using sequence alignment techniques.

There are several works related to activity modeling from location-driven phone sensor data. CitySense [18] is a mobile application which uses GPS and WiFi data to summarize "hotspots" of activity in the San Francisco area, which can then be used to make recommendations to people regarding, for example, preferred restaurants and nightclubs [23]. Liao *et al.* [17] use GPS data traces to label and extract a person's activities and significant places. Their method is based on Relational Markov Networks. Eagle and Pentland [7], the pioneers in the Reality Mining research domain, used principle component analysis (PCA) to identify the main components structuring daily human behavior. The main components of human activities, which are the top eigenvectors of the PCA decomposition are termed *eigenbehaviors*. Our previous work [8] builds on the initial ideas in [7], though we propose the use of probabilistic topic models and develop flexible feature bags to capture human routines in a robust manner (i.e., small variations in daily activities will not affect results though they might result in eigenbehavior changes). Our method also had the advantage of capturing characteristic trends occurring over part of the day (such as early morning only), whereas eigenbehaviors capture features over the entire day.

To our knowledge, relatively few works have focused on large-scale human activity modeling from proximity or multimodal mobile sensor data. There are some works by Pentland's group [20], [21], using multimodal data, though the critical features for these works are nonverbal audio features which would not be readily available from most off-the-shelf mobile devices. In [6], a dynamic proximity network is modeled and analyzed to find the properties of human interaction dynamics. In [24], a study of how mobile phone viruses spread investigated joint location and proximity mobile phone data, though the focus of that work is to the application of epidemiology rather than to the mining of routines as we do here. Recently, we did a preliminary study on Reality Mining data, where we investigated human activity patterns from multimodal mobile data, considering both location and proximity data [9]. This paper extends that initial work by providing further details and analysis of the data and results. More specifically, this paper introduces the concept and methodology in more details. Further, we introduce a detailed analysis of user interactions within the group and with other Bluetooth devices and compare the interactions of two different subpopulations. We also extend the multimodal routine discovery to consider factors such as the day of the week, leading to the discovery of work patterns dominating on Sundays versus Mondays, for example. We also present an investigation of user entropy to differentiate varying types of individual behaviors.

III. Multimodal Framework

We use the Reality Mining (RM) dataset [7] for which the activities of 97 students and staff at MIT were recorded by Nokia 6600 smart phones over the 2004–2005 academic year. Given

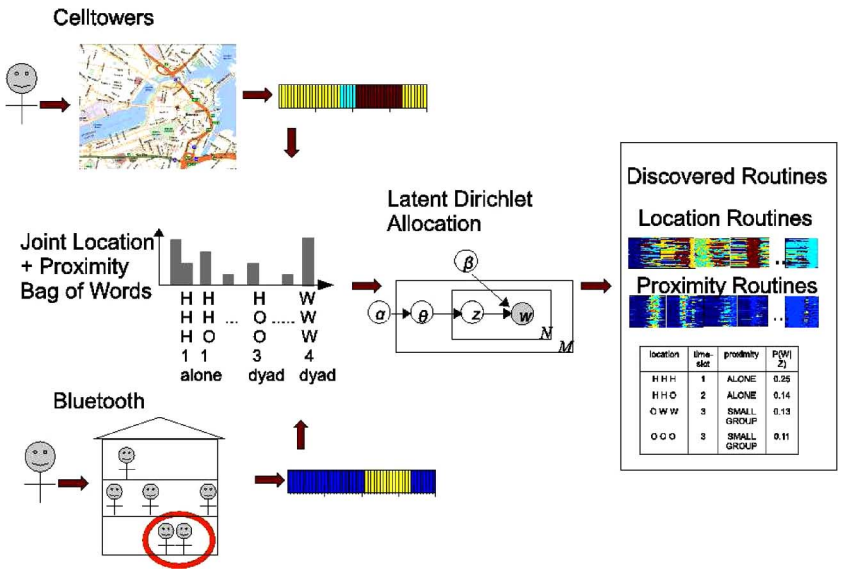

Fig. 1. Overview diagram of our method. The data captured by mobile phones (where a user is as well as with whom) is combined to form a joint location-proximity representation. After the multimodal data representation is transformed to a bag of words, Latent Dirichlet Allocation inference is applied to reveal latent topics (or discovered routines), corresponding to common user places and interactions. Each routine is characterized by its top multimodal words ranked by their probability.

a day in the life of a person in terms of where they go and the number of people within the group they are in proximity with, our goal is to discover routines from large-scale multimodal phone data. Further, we use the combined location and proximity routines discovered to predict missing location and proximity data. An overview of our method is visualized in Fig. 1. We represent a day in the life of a user in terms of where they are over a 90-min time interval as well as the number of people they are with during this time interval within the RM population, forming a joint location-proximity data representation, described next. This joint data representation is input to the Latent Dirichlet Allocation (LDA) model, from which human routines are discovered, representing common forms of social interactions which occur at varying locations.

A. Joint Location-Proximity Representation

The joint location-proximity data representation is based on the concatenation of data corresponding to users' location, proximity, and a timeslot indicating a coarse-grain measure of the time of day for which this data is measured. The details follow.

Location Representation: Following Eagle *et al.* [7], a given individual's locations (given by cell towers) is represented over the course of a day by representing all possible locations into four categories, namely work (W), home (H), out (O), and no reception (N). W are the MIT work premises, H are the homes of individuals, and O are towers that are not H or W, thus encompassing a large number of places. The W and H labels can be, in general, easily obtained from user tagging of cell towers or from knowledge about the data collection campaign. N is a label used if there is missing data for a person for a given time, for instance when the phone is off. The basic idea for the location representation, which is taken from our previous work [8], is to assign a single location label (H, W, O, N) for each 30-minute time interval of a user's day, resulting in 48 location labels for each user and each day. The use of 30-minute slots, synchronized on the hour, is a simple yet robust assumption, as many people and organizations schedule their life around this type of day segmentation. Also, the data is quite noisy and challenging, and this representation aids with some sources of noise, such as the numerous fluctuating cell tower recordings. To assign a single label to a 30-minute slot, we compute the time of occurrence for all location labels within the slot, and assign the one with largest duration. Then, three consecutive 30-minute labels are taken to obtain location transition information over a 1.5 hour period in a day. These 1.5-hour intervals are overlapping, resulting in 48×1.5-hour 3-label location sequences in a day. We use 1.5-hour intervals in order to capture transitions in user movement.

Proximity Representation: For proximity data, we use the Bluetooth readings to consider proximity with people in the Reality Mining group. Bluetooth can detect other similar devices located within a 10-m radius. Bluetooth is a reasonable (although clearly imperfect) proxy for social interactions, though there are various sources of noise making it challenging to work with. On one hand, we could expect that people actually interacting will often be sensed by Bluetooth but many cases of nearby people who do not interact will be detected too. This is a limitation of the Bluetooth modality. Proximity in general could be considered as proximity to laptops, computers, and other people, is also recorded in the data, but it is difficult to distinguish them from mobile phones. We quantize the number of proximate people into four prototypical groups: user is alone, dyad (one person in proximity), small group (two–four people in proximity), large group (five or more people in proximity). The group sizes are motivated by research in social science that has traditionally analyzed dyads, small groups, and large groups as separate categories, as they present distinct dynamics [10].

Timeslot Division: Each day is divided into eight coarse-grain 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), 9–12 pm (8). These timeslots were chosen to capture common events

// GOAL: Given a training corpus and parameters α, β, and T, estimate the parameters $n_d^{(t)}$ and $n_t^{(w)}$ from which we can determine the model parameters $\hat{\phi}_t^{(w)}$ and $\hat{\theta}_d^{(t)}$.

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

// 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 $n_d^{(t)} = n_d^{(t)} - 1$, $n_t^{(w)} = n_t^{(w)} - 1$.
    8) Sample a topic index t from $p(z_i = t|\mathbf{z}_{\neg i}, \mathbf{w}) \propto \frac{n_{t,\neg i}^{(w)} + \beta}{\sum_{w=1}^{V} n_{t,\neg i}^{(w)} + \beta} \cdot \frac{n_{d,\neg i}^{(t)} + \alpha}{\sum_{t=1}^{T} n_d^{(t)} + \alpha}$.
    9) Increment the new word/topic and topic/document counts as follows $n_d^{(t)} = n_d^{(t)} + 1$, $n_t^{(w)} = n_t^{(w)} + 1$.

// Compute model parameters
10) Determine the unknown parameters as follows
$\hat{\phi}_t^{(w)} = \frac{n_t^{(w)} + \beta}{n_t + V\beta}$, and $\hat{\theta}_d^{(t)} = \frac{n_d^{(t)} + \alpha}{n_d + T\alpha}$, where $\hat{\phi}$ and $\hat{\theta}$ are the model parameter estimates, $n_t = \sum_{w=1}^{V} n_t^{(w)}$, and $n_d = \sum_{t=1}^{T} n_d^{(t)}$.

Fig. 2. Gibbs Sampling algorithm for LDA.

in daily life, such as lunch time, dinner time, or morning and afternoon work times. Other time intervals could equally be used to capture events occurring over finer or coarser daily periods.

A day in a user's life is finally represented as a *multimodal bag of words*, where a word is a location sequence, concatenated with the corresponding proximity group and a timeslot, as shown in Fig. 1. The bag of word model is amenable for probabilistic topic modeling which is introduced in the next subsection.

B. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an unsupervised probabilistic generative model that was initially developed to characterize text collections, but can be extended to other collections of discrete data [2]. A *word* is a basic unit of discrete data defined by an item for a vocabulary of size V. A *document* is a bag of N words, and a corpus is a collection of M documents. Each document is viewed as a mixture of topics, where topics are distributions over words. The probability of a word w in a document, assuming it is generated from a convex combination of T topics, is given as

$$p(w) = \sum_{t=1}^{T} p(w|z = t)p(z = t) \tag{1}$$

where z is a latent variable indicating the topics from which the word w was drawn. In LDA, a Dirichlet prior is assumed on the topic distributions to provide a complete generative model for documents [12]. The graphical model for LDA is shown in Fig. 1.

The objective of LDA inference is to determine the word distribution $p(w|z = t) \stackrel{\text{def}}{=} \phi_t^{(w)}$ for each topic t, and the topic distribution $p(z = t) \stackrel{\text{def}}{=} \theta_d^{(t)}$ for each document d. We use the approximation derived in [12] based on Gibbs sampling. In LDA, $p(\theta)$ and $p(\phi)$ are assumed to be Dirichlet distributions with hyperparameters α and β, respectively. The Gibbs sampler is used since exact inference is intractable [12]. Let $n_t^{(w)}$ and $n_d^{(t)}$ be the number of times word w and document d have been assigned to topic t, respectively. Let $n_t = \sum_{w=1}^{V} n_t^{(w)}$ and $n_d = \sum_{t=1}^{T} n_d^{(t)}$ denote the sums of words in a given topic, and of topics in a given document, respectively. Let $\mathbf{w}$ denote the set of words in the corpus, and $\mathbf{z}$ denote the set of topics in the corpus, and $\mathbf{z}_{\neg i}$ denote $\mathbf{z}$ excluding the current topic element $z_i = t$. In practice in the Gibbs sampler, we sample from

$$p(z_i = t|\mathbf{z}_{\neg i}, \mathbf{w}) = \frac{p(\mathbf{w}, \mathbf{z})}{p(\mathbf{w}, \mathbf{z}_{\neg i})} \tag{2}$$

$$= \frac{p(\mathbf{w}|\mathbf{z})}{p(\mathbf{w}_{\neg i}|\mathbf{z}_{\neg i})p(w_i)} \cdot \frac{p(\mathbf{z})}{p(\mathbf{z}_{\neg i})} \tag{3}$$

$$\propto \frac{n_{t,\neg i}^{(w)} + \beta}{\sum_{w=1}^{V} n_{t,\neg i}^{(w)} + \beta} \cdot \frac{n_{d,\neg i}^{(t)} + \alpha}{\sum_{t=1}^{T} n_d^{(t)} + \alpha} \tag{4}$$

using the procedure summarized in Fig. 2. In the above equations, $n_{.,\neg i}^{(.)}$ denotes the counts of the elements jointly contained in the subscript and superscript, excluding the current element i. The topic assignments, $n_d^{(t)}$ and $n_t^{(w)}$ are initialized randomly. In each Gibbs sampling iteration, the topic assignments for a word and a document are sampled from (4). After a predefined number of iterations (i.e., after the burn-in time of the Gibbs sampler), the sampler is assumed to have approached its stationary distribution [22]. This is a common assumption in MCMC methods. Essentially, the initialization process randomly assigns words and documents to topics. Then the chain is run in order to "refine" these assignments according to (4). In this paper, we use the last sample in the chain for document and word ranking of topics due to the lack of identifiability problem

// GOAL: Given a test document $\tilde{d}$ with missing location and proximity labels for timeslot s_m, predict a label.

// Topic discovery from the training corpus.
1) The Gibbs sampling algorithm in Figure 2 is performed on the training corpus to discover topics.

// LDA querying is performed to retrieve documents relevant to test documents.
2) Follow the Gibbs sampling procedure in Figure 2, replacing the topic sampling in Step 5 by the following equation, from which topic t is sampled:

$$p(z_i = t|\mathbf{z}_{\neg i}, \mathbf{w}) \propto \frac{n_t^{(w)} + \tilde{n}_{t,\neg i}^{(w)} + \beta}{\sum_{w=1}^{V} n_t^{(w)} + \tilde{n}_{t,\neg i}^{(w)} + \beta} \cdot \frac{n_{\tilde{d}\neg i}^{(t)} + \alpha}{\sum_{t=1}^{T} n_{\tilde{d}}^{(t)} + \alpha}, \tag{6}$$

where $\tilde{d}$ represents the test document, and $\tilde{n}_{t,\neg i}^{(w)}$ counts the observations of term w and topic t in the test documents, excluding the i^{th} index [13].

3) The document/topic distribution for the test document $\tilde{d}$ is $\hat{\theta}_{\tilde{d}}^{(t)} = \dfrac{n_{\tilde{d}}^{(t)} + \alpha}{n_{\tilde{d}} + T\alpha}$.

// Find the best matching topic for test document $\tilde{d}$.
4) The topic z_i for which $i = \underset{j}{\text{argmin}}|s_j - s_m|$ and $\hat{\theta}_{\tilde{d}}^{(z_i)} > Th$, where s_i is the timeslot of the most probable word of topic z_i, s_m is the timeslot of the missing data, $s_i, s_m = \{0, ..8\}$, and Th is a threshold, is chosen as $z_{\tilde{d}}^{top}$.

//Replace the missing data.
5) $\tilde{d}(s_m) = d_{z_{\tilde{d}}^{top}}(s_m)$, where $d_{z_{\tilde{d}}^{top}}$ is the most probable document given $z_{\tilde{d}}^{top}$.

Fig. 3. Algorithm for predicting proximity and location timeslots.

in sampling-based LDA (i.e., there is no guarantee that topics across samples are the same [22]). The Gibbs sampler results in

$$\hat{\phi}_t^{(w)} = \frac{n_t^{(w)} + \beta}{n_t + V\beta}, \quad \hat{\theta}_d^{(t)} = \frac{n_d^{(t)} + \alpha}{n_d + T\alpha}. \tag{5}$$

In our work, documents are days in people's lives and words are the location-proximity words defined in Section III-A. Topics are expected to correspond to routines.

In this work, we use LDA for two tasks:

Routine Discovery: We propose to extend the use of LDA to handle multimodal data, expecting that topics will capture joint patterns of location and proximity that help disambiguate relevant cases (e.g., discriminating between a person at work alone and in a group). Routines can be identified by observing the top words for a given topic (ranked by their probability) and also by the top days for a given topic.

Predicting Behavior: LDA is also used for the prediction of missing labels in a day (i.e., the prediction of users' joint patterns of location and proximity for certain timeslots). To achieve prediction, LDA inference is run on the test days containing missing bits. The algorithm details are presented in Fig. 3. s_m is defined as the timeslot of a document (a day), where $m = 1 \ldots 8$ are the eight coarse-grain possibilities in a day. After finding topics within the training corpus via LDA, a distribution of topics for each test document, $\tilde{d}$, is inferred resulting in $\hat{\theta}_{\tilde{d}}^{(t)}$. The resulting topics for document $\tilde{d}$ are ranked according to $\hat{\theta}_{\tilde{d}}^{(t)}$ and the best matching topic for document $\tilde{d}$ is denoted by $z_{\tilde{d}}^{\text{top}}$, which is found according to Step 4 in Fig. 3. The result is a single topic which is used for replacement of the missing data over the timeslot. To fill in the missing location and proximity words, we replace the missing labels with those of the top day for the mostly likely topic selected; $\tilde{d}(s_m) = d_{z_{\tilde{d}}^{\text{top}}}(s_m)$, where $d_{z_{\tilde{d}}^{\text{top}}}$ is the most probable document given $z_{\tilde{d}}^{\text{top}}$. For the predicting behavior task (whose results are given in Section IV-D), experiments are performed over ten chains. Note that the procedure used for behavior prediction described here is simple and more elaborate methods to predict missing labels could be derived from the output generated by LDA.

IV. Experiments and Results

A. Data and Model Parameters

We experimented with all of the 97 individuals in the RM dataset and with days ranging from 18.07.2004 to 05.05.2005, encompassing 291 consecutive days thus extending our previous work [8] which only considered 30 users. This subset of days was chosen since these are the days for which proximity data is mostly available. Days with entirely no reception for location were not considered since they contain no useful information for proximity either. The LDA model for joint location-proximity routine discovery used $T = 100$ topics. Heuristic methods were used to obtain T, but generally speaking, a small value of T will produce coarse routines, whereas a large T will be much more specialized. The estimation of the optimal number of topics in topic models is an active research problem [1], [13]. The hyperparameters were set to $\beta = 0.01$ and $\alpha = 50/T$. These hyperparameters are chosen based on standard values used for text analysis [12].

B. Exploratory Analysis

We performed an analysis of the proximity data to study the interactions of business students compared to engineering students and staff, considering interactions for different days of the week as well as interactions with others in the same group compared to other Bluetooth devices (not including people in the group), which could include family members, friends, strangers, laptops, or computers. The results are illustrated in Fig. 4. The

Topic 11 Top 3 words	$p(w\|z)$
W W W 11am-2pm Small Group	0.502
W W W 2pm-5pm Small Group	0.427
H H W 9am-11am Dyad	0.015

Topic 28 Top 3 words	$p(w\|z)$
O O O 9pm-12pm Alone	0.634
O O O 7pm-9pm Alone	0.358
W O W 11am-2pm Dyad	0.001

Topic 41 Top 3 words	$p(w\|z)$
W W W 2pm-5pm Large Group	0.610
W W W 5pm-7pm Large Group	0.275
W W H 5pm-7pm Large Group	0.034

Topic 46 Top 3 words	$p(w\|z)$
W W W 7pm-9pm Small Group	0.415
W W W 9pm-12pm Small Group	0.219
W W W 5pm-7pm Small Group	0.132

Topic 53 Top 3 words	$p(w\|z)$
H H H 0-7am Alone	0.757
H H H 7am-9am Alone	0.215
H H W 9am-11am Alone	0.015

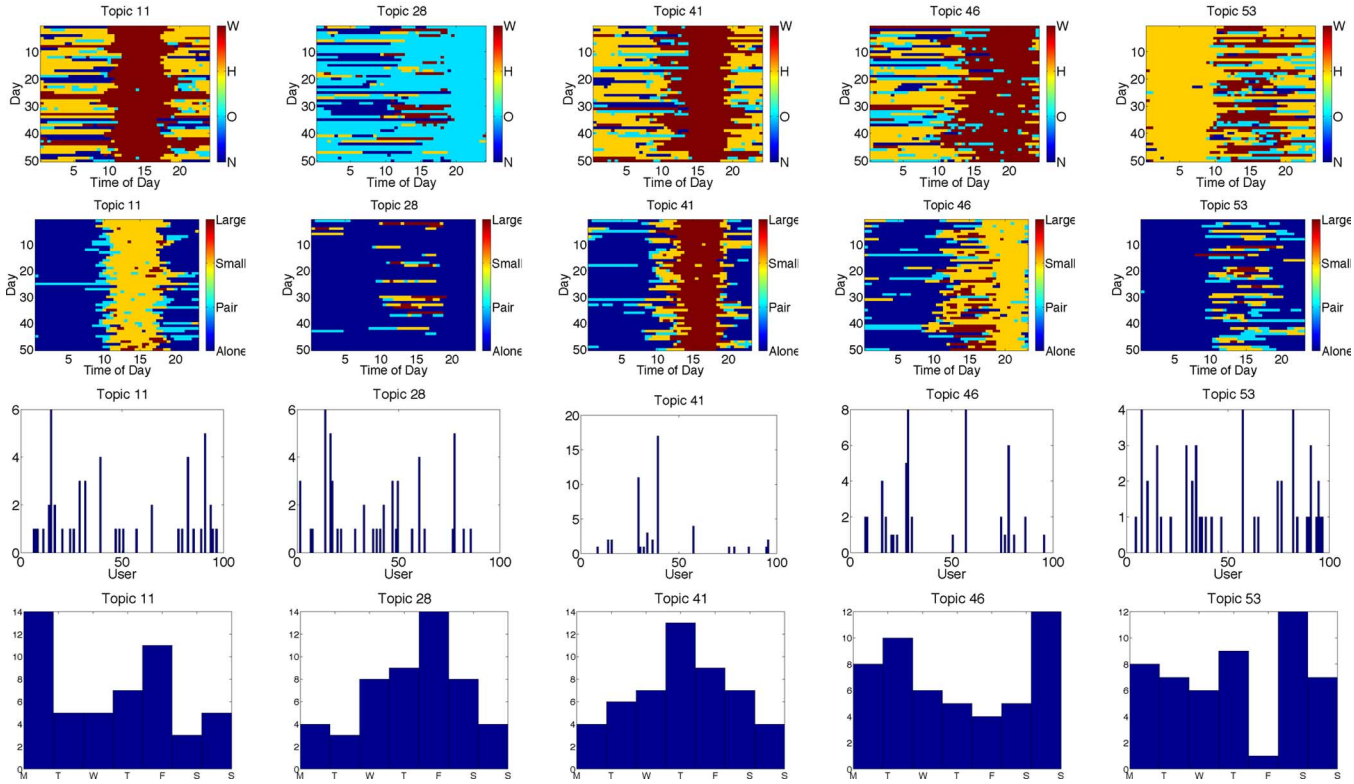

Fig. 4. Interaction patterns of MIT business students compared to engineering students and staff. (a) and (b) visualize the pairwise user interactions in terms of (a) the number of interactions and (b) the total duration of interactions (hours). Business students (1–27) and Media Lab users (28–97) are highlighted by boxes. There are many interactions between engineering students which do not occur over long durations. The average quantity of interactions over all Sloan business students versus all Media lab students and staff is computed over the days of the week "S M T W T F S" in terms of (c) the number of interactions and (d) the total duration of interactions (hours). On average, Media Lab users have more interactions, though on Thursdays business students interact for longer durations, perhaps due to a course on this day. They also interact less on Mondays, Wednesdays, and Fridays. On average, there is little interaction on weekends in all cases. The total interaction times (hours) of users with other Reality Mining users in comparison to all other Bluetooth devices are shown in (e) for Sloan students and (f) for Media Lab users.

entire Reality Mining dataset was considered for these results, including 16 months of 97 users' data.

Fig. 4(a) and (b) illustrates the quantity of interactions between users of the Reality Mining study. Users 1–27 are the Sloan business students, and users 28–97 are the Media lab students and staff. There are two boxes marking the separation between those groups in Fig. 4(a) and (b). We plot the quantity of interactions between individuals in terms of (a) the number of interactions during the course of the study (without taking into account the duration of interaction) as well as (b) the total duration of interaction between these users in hours. In both plots, the amount of interaction (either considering number of interactions or total duration) was much higher between several Media Lab users, in comparison to business students. The figures have been adjusted to visualize the interaction between business students as well by assigning any interactions occurring over a threshold to the last bin of the colorbar (200+ interactions or 150+ hours). More specifically, in Fig. 4(a), if there are 200 or more interactions between a pair of users, this is labeled by 200+. The threshold 200 is chosen by rounding up the maximum number of interactions between business students. The same procedure is applied in Fig. 4(b) for hours of interaction. The maximum number of interactions throughout the study occurred between a pair of Media Lab users, and was approximately 585. The maximum duration of interactions occurred between a differing pair of Media Lab users, and was on the order of 690 hours over the course of 16 months. Note that these plots are not symmetric due to the inconsistencies in Bluetooth and the data collection software. Often times, two people will be sensed as being proximate only by one of the phones. Furthermore, there are several users without any data recordings. There are many interactions which occur frequently between individuals though not for long durations. This is especially visible between several of the Media Lab users. Also note that in-

teractions between business students and engineering students are quite sparse. There are many Media Lab users that never interact, though most business students (with data recorded) interact, resulting in a much less sparse matrix. There was a pair of users with negative duration values, likely due to incorrect clock settings, which was removed.

Fig. 4(c) and (d) plots the overall means of the number of interactions and the duration of interactions in hours respectively, for Media Lab and business users over the week where "S M T W T F S" on the x-axis corresponds to "Sunday through Saturday." These average values varied greatly across users. We can see in both groups of people, the interactions are very low on the weekends. The mean number of interactions is always on average higher for Media Lab students on every day of the week, though it is especially higher on Mondays, Wednesdays, and Fridays. The duration of interactions for Sloan students is on average higher than Media Lab students on Thursdays, perhaps due to a course or business school event on this day.

In Fig. 4(e) and (f), we plot the total duration of user interactions with users in the study compared to "non-user" Bluetooth devices (or other devices), which could include family, friends, strangers, laptops, and computers. Fig. 4(e) illustrates the total interaction times of Sloan users whereas (f) is for the Media Lab users. In (e) and (f), we can see there are a few people in both groups who have heavy interactions within the group. Also, many of the users have more interaction with people in the group than with "other devices." Many of the Media Lab users have heavy interactions with "other devices," likely due to the fact that they spend hours in front of their laptops and computers daily.

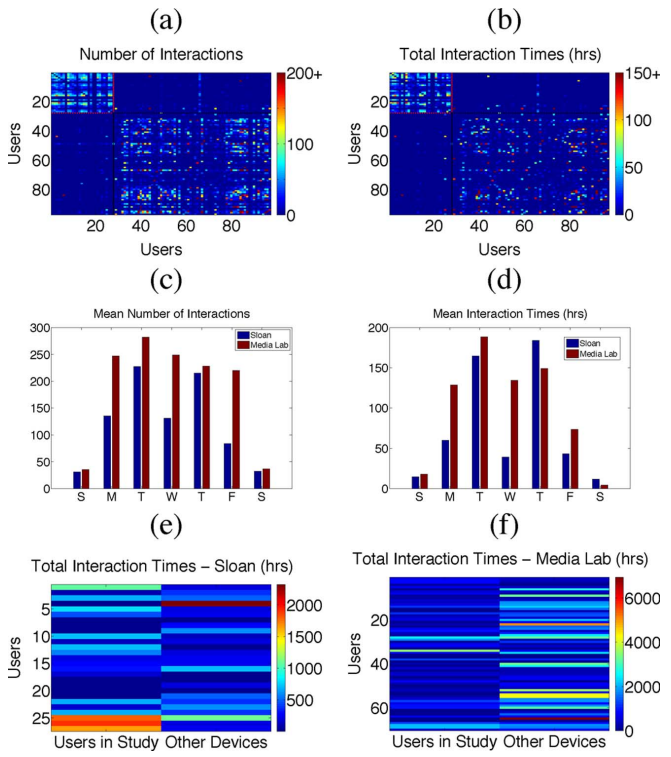

Fig. 5. Selected LDA results. The first row of tables correspond to the most probable words given a topic. Ranked days (i.e., documents) for selected topics by $p(d|z_i)$, showing (second row) the top 50 days' location data and (third row) the corresponding proximity data for a given topic. (fourth row) Histograms of the users whose days ranked in the top 50 for topic z_i. (last row) Histograms of the days of the week (M T W T F S S = Monday to Sunday) that ranked in the top 50 for topic z_i. Note the colorbars for the location figures indicating the W, H, O, and N locations, and for the proximity figures indicating a large group, small group, pair, or alone.

C. Joint Location-Proximity Routines

The fusion of proximity and location data enables the discovery of more detailed patterns regarding this group of MIT users' daily lives compared to single modalities. After LDA learning, there is a chance that two topics could be similar to each other, as LDA does not guarantee that topics be distinct from each other. The fact that LDA-learned topics are often similar to each other has also been observed in the text domain. A short summary of the learned routines on the entire corpus is presented next, and a summary is visualized in Fig. 5.

- *Home routines and proximity:* Most of the home routines discovered occurred for users alone (i.e., not in proximity with anyone from the group). Only 2 out of the 20 topics related to discovered home routines dominated for a pair of users in proximity. No home routines occurred for small or large groups in proximity, which suggests that people did not socialize within the population at home.
- *Work routines and proximity:* Most of the routines discovered with proximity interactions occurred at work locations. There are 17 topics corresponding to work routines, and 13 of them occur with proximity patterns. Routines at work were discovered for all four proximity groups (users alone, in dyads, small, and large groups), which indicates that all these types of interactions occur frequently.
- *Morning routines and proximity:* Only 3 out of 100 topics had a proximity interaction in the morning (before 10 am), and all 3 of these routines occur for pairs of users and never for groups. People interacting in the morning seems to be relatively sparse for this population.
- *Day time routines and proximity:* Approximately 20 topics characterize user interactions throughout the day (10 am–7 pm). The interactions include pairs of users, as well as small and large groups.
- *Evening routines and proximity:* 7 topics characterize group interactions in the evenings (7 pm–midnight). These occur for pairs of users, and small as well as large groups.

A selection of topics illustrating the types of joint routines discovered are visualized in Fig. 5. We have illustrated results for selected topics, $z_i = 11, 28, 41, 46, 53$, for the 50 most probable days given those topics. The three most probable words given the topics are shown in the tables in the first row. We plot the results in terms of users' locations (second row), proximity (third row), user statistics (fourth row), and day of week statistics (fifth row). The figures illustrating the users' locations and proximity data show the time of the day as the x-axis, and each row is a day of the life of a user plot in terms of their location (H is home, W is work, O is out, and N is no reception) as well as in terms of their interactions where ("large" corresponds to a large group, and "small" to a small group). Furthermore, a histogram for the users whose days ranked in the

top 50 documents is shown in the fourth row, the x-axis indicating anonymous user id and the y-axis the number of days. The fifth row illustrates a histogram of the days of the week (M T W T F S S = Monday to Sunday) of the 50 most probable days given each topic. A summary of the routines discovered plotted in Fig. 5 is as follows:

- **Topic 11**: The user is at work during the day (dominantly 11 am–5 pm as seen from the three top words given topic 11) while in proximity with a small group of 3–5 people. Several users have days with high probability of topic 11. This work routine dominates on Mondays.
- **Topic 28**: The user is out in the evenings (7 pm–12 pm) alone. This routine occurs most frequently on Fridays for several users in the study.
- **Topic 41**: The user is at work from 2 pm–7 pm in a large group. This occurs dominantly for a handful of users, predominantly on Thursdays. Note, that most of these users correspond to Sloan business school students, displaying their common Thursday afternoon work routine.
- **Topic 46**: The user is at work in the evening (from 5 pm–midnight) in a small group. This work routine dominates on Sundays and occurs often for a few users.
- **Topic 53**: The user is at home alone in the mornings (from midnight until 11 am). This topic hardly ever occurs on Fridays.

D. Behavior Prediction

We now show how it is possible to use LDA in order to predict unobserved location and proximity data for a timeslot of a user's day. For experiments, we decided to distinguish between people based on the entropy of their routines under the hypothesis that prediction of location and proximity will be more or less difficult depending on the variability of each person's habits. User entropy is computed on the distribution of topics given users, $p(z|u) = \sum_d p(z|d,u)p(d|u)$, where u is the user variable, and we assume $p(d|u) = (1/|D_u|)$, D_u is the set of users recorded for user u, and $|D_u|$ is the set cardinality. The topics z correspond to the joint location-proximity routines found in Section IV-C. All of the users in the dataset are ranked according to their entropy. After this, we set two thresholds for high and low entropy which gave ten users in each case. We randomly picked five people for each class (high and low entropy).

For each of the ten selected users, 20 days of their life were randomly selected from days with at least one proximity interaction (i.e., days that contained at least one non-empty word over the entire day). This set of days was used to form the test set, from which we systematically remove words to generate data with missing sequences to predict. For each day, the words of a given coarse-grain timeslot were removed to form a day for which the method has to predict the missing sequence, thus generating 8 days, each with one timeslot's words missing. The resulting dataset for which we predict missing sequences contains ten users, each with 160 days = 1600 documents for testing. Thus, for each user there are 160 documents for testing, and each coarse-grain timeslot contains 200 documents for testing.

For each document, there is one timeslot with missing location and proximity labels. For evaluation, we compute two types

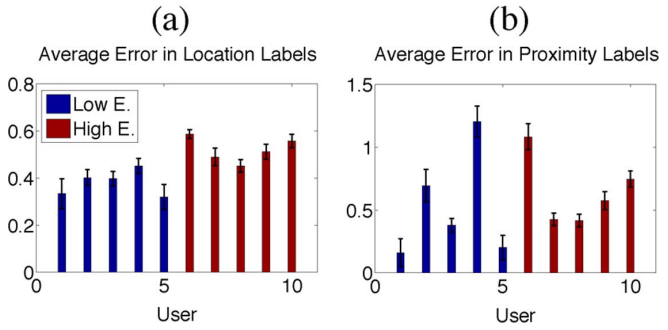

Fig. 6. (a) Average location prediction error as a function of users, where low entropy users are labeled "Low E" and high entropy users "High E." (b) Average proximity prediction error as a function of users. Location label for prediction is consistently lower for low entropy users. However, for proximity errors are not necessarily lower for low entropy users.

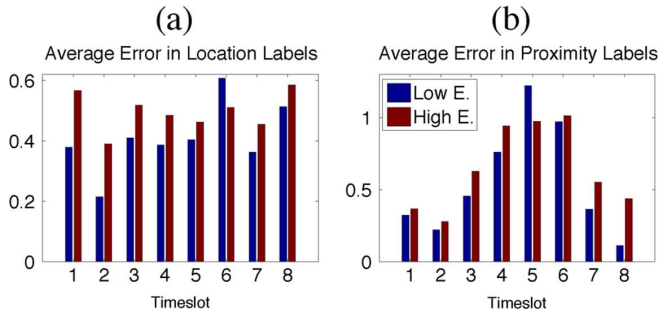

Fig. 7. Average error in (a) location prediction, and (b) proximity prediction, as a function of timeslot for low and high entropy users. High entropy users consistently have higher location label errors for prediction over all times of the day, though the error is highest between 5–7 pm (timeslot 6) which corresponds to typical commuting times. The highest errors in proximity label prediction occur from 9 am–7 pm, corresponding to work times where most interactions occur.

of error. The *location error* is the number of incorrectly predicted location labels divided by the total number of labels to be predicted in the given coarse-grain timeslot. For instance, documents with timeslot 1 missing have 14 location labels to be predicted since it occurs from 0–7 am. The *proximity error* is the average number of people wrongly predicted for each word in a given timeslot. More specifically, if the predicted group (alone, dyad, small group, large group) is correct then there is no error. If the predicted group is incorrect, then we predict the minimum number of possible people in the group (alone = 1, dyad = 2, small group = 3, large group = 5) and compute the difference with the actual number of people in proximity. For example, if there are ten people in proximity and we predict a small group, then we assume three people are in proximity. If this incorrect prediction occurs over the 14 half-hour words in timeslot 1 (midnight–7 am), then the average proximity error is 7. Finally, the results for location and proximity error are averaged over ten randomly initialized chains of the Gibbs sampling procedure described in Fig. 2.

The location and proximity errors are computed over users and timeslots and displayed in Figs. 6 and 7. We present the average errors as a function of the user for location in Fig. 6(a) and for proximity in Fig. 6(b). Users 1–5 (in blue) have low entropy and 6–10 (in red) have high entropy. Interestingly, low-entropy users have lower error in the prediction of location labels than high-entropy users. For low entropy users, the error can be as

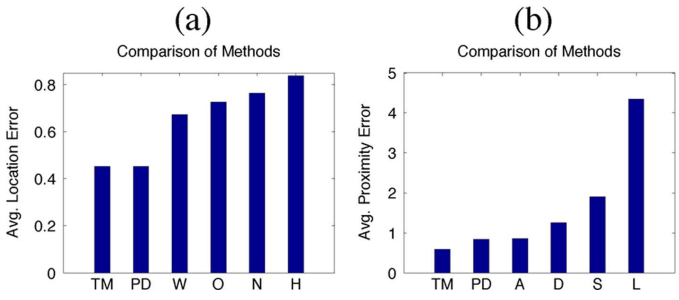

Fig. 8. Comparison of our topic model (TM) approach to various other methods for overall location and proximity errors. PD is the nearest neighbor approach of replacing data with the previous days'. (a) For the overall average location error W represents the error obtained if all missing data is replaced by work, O by other, N by no reception, and H by home. (b) For the overall average proximity error A represents the error obtained if the missing data is replaced by alone, D by dyad, S by small group, and L by large group. The TM approach predicts missing location data as well as the PD approach; however, our approach outperforms the PD method for predicting missing proximity data. The TM also outperforms all the other methods (both in terms of location and proximity missing data prediction) significantly.

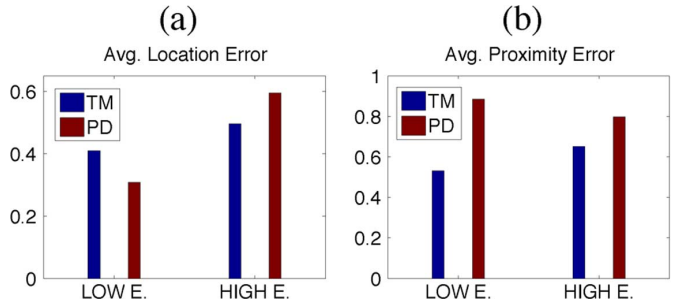

Fig. 9. Comparison of our topic model (TM) approach with the previous day (PD) approach in terms of user types. (a) Average location error for low entropy users and high-entropy users for the TM versus PD approach. The PD approach performs better for location data prediction for low entropy users; however, our TM approach performs better for high entropy users. (b) Average proximity error for low and high entropy users for the TM versus PD approach. In both cases, our TM approach outperforms the PD approach for proximity data prediction.

low as 0.32 which nevertheless indicates that the task is difficult. We also include errorbars corresponding to the standard deviation over the ten randomly initialized chains. High entropy users are significantly more difficult to predict. In Fig. 6(b), we plot the proximity error. In the best (resp. worst) case, the predicted number of people in proximity is incorrect by, on average, 0.16 (resp. 1.2) people. In this case, low entropy users do not necessarily have lower prediction errors in proximity than high entropy users.

In Fig. 7(a) and (b), we plot the average errors as a function of coarse-grain timeslot for both high and low entropy users for location [Fig. 7(a)] and proximity [Fig. 7(b)]. We can see in Fig. 7(a) that for almost every timeslot (with the exception of timeslot 6), high entropy users are harder to predict (have higher errors) than low entropy users. Timeslot 6 (5–7 pm, which corresponds to typical commuting times) is overall the most difficult to predict. Also, for timeslots 1 and 2 (midnight to 9 am), low entropy users correspond to much better performance than high entropy users. Regarding Fig. 7(b), the error in proximity prediction as a function of timeslot is again not highly correlated with the entropy of a user. The prediction in proximity has the highest error in timeslot 5, corresponding to 2–5 pm, and the lowest error in the mornings and late evenings, which is not surprising. In the worst case, the proximity error in any given timeslot is less than 1.25 people on average.

In Fig. 8, we compare the performance of our topic model (TM) method to several other methods. Fig. 8(a) illustrates the overall average location error for the TM approach in comparison to a nearest-neighbor approach called previous day (PD), which uses knowledge about the specific date of the test day, and replaces the missing data with that of the previous day. Note that the date is a very strong contextual cue about human routines that is not currently used in our method TM. The approach labeled W is the case where all missing data is replaced by the "work" location. Similarly, O is the case where all missing data is replaced by "out," N by "no reception." and H by "home." Fig. 8(b) illustrates the overall average proximity error for the TM approach in comparison to the PD approach, in addition to the approaches labeled A, D, S, and L, corresponding to replacing the missing proximity data with the labels "alone," "dyad," "small group," and "large group," respectively. We can see the TM and PD approaches perform similarly in terms of location data prediction, however, TM outperforms PD for missing proximity data prediction. The TM approach also outperforms all the other methods illustrated.

Given the simplicity of the PD method, we look deeper into the TM and PD performance for various types of users in Fig. 9. Fig. 9(a) illustrates the average location prediction error for high and low entropy users. We see that for location prediction, the PD method performs better for low entropy users. This is understandable since low entropy users have very "routine" lifestyles and simply replacing the missing data with that of the previous day results in good performance. However, for high entropy users, our TM method, which captures specific patterns of transitions (e.g., H to W), is working better. Given these complimentary features, for future work, we plan to investigate a method that integrates both concepts. Fig. 9(b) illustrates the average proximity prediction error for high and low entropy users. The results show that our TM approach outperforms the PD approach both for low- and high-entropy users.

V. Conclusion

We have proposed a probabilistic methodology that successfully discovers recurrent patterns in people's lives from multimodal data, and that can use the discovered routines for data prediction, estimating location, and proximity data of users with varying entropy. Essentially, the method mines the most dominantly occurring human routines (topics) from a huge corpus of real-life human mobile data to determine recurring human patterns involving time of the day, semantic location, and proximity based interaction type. Our method also uses these rich human location-interaction topics to predict missing data, which in real life occurs very frequently with mobile phone data, and can also be seen as a method to verify the validity of the routines discovered. By computing the entropy of individuals based on their jointly modeled locations and interactions, our method is able to predict missing multimodal data over several hours for users with both low and highly varying lifestyles.

In future work, the methodology for data prediction could be further optimized to use the topics in a more sophisticated

manner, and to include prediction on varying timescales, such as full days of missing data. It would also be very useful to take advantage of the other, often available data modalities of mobile sensor data for data prediction. For instance, one could predict a user's location given the time of day and their interactions, the day of the week, or even using their phone call and SMS data. The Bluetooth proximity data is potentially a very rich source if one considers proximity to all other devices including laptops, computers, and anonymous cell phones. This data in itself could be used to determine the semantic labels of an individual, such as if the user is at home (in proximity with their home computer), at work (in proximity with their work computer), or out (in proximity with strangers). In a different line of work, we would like to enrich the location vocabulary by refining the "other" category. This in principle could be done from the Reality Mining dataset, but handling sparse human annotation of places is in itself a research problem. Finally, we would like to consider recent approaches like the Maximum-Margin Supervised Topic Model [25] which explicitly addresses the issue of maximizing the distance between topics, and may be used to optimize the number of topics output.

ACKNOWLEDGMENT

The authors would like to thank N. Eagle (Santa Fe Institute) and A. (Sandy) Pentland (MIT) for sharing the data.

REFERENCES

[1] D. Blei and J. Lafferty, *Text Mining: Theory and Applications*. New York: Taylor & Francis, 2009, "Topic models," .
[2] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 4–5, pp. 993–1022, 2003.
[3] J. Candia, M. Gonzalez, P. Wang, T. Schoenharl, G. Madey, and A. Barabasi, "Uncovering individual and collective human dynamics from mobile phone records," *J. Phys. A: Math. Theoret.*, vol. 41, no. 22, 2008, 224015.
[4] T. Choudhury, M. Philipose, D. Wyatt, and J. Lester, "Towards activity databases: Using sensors and statistical models to summarize people's lives," *IEEE Data Eng. Bull.*, pp. 49–58, 2006.
[5] D. Choujaa and N. Dulay, "Activity inference through sequence alignment," *Location Context Awareness*, pp. 19–36, 2009.
[6] A. Clauset and N. Eagle, "Persistence and periodicity in a dynamic proximity network," in *Proc. DIMACS Workshop Comput. Methods Dyn. Interact. Netw.*, 2007.
[7] N. Eagle and A. Pentland, "Eigenbehaviors: Identifying structure in routine," *Behavioral Ecol. Sociobiol.*, vol. 63, no. 7, pp. 1057–1066, 2009.
[8] K. Farrahi and D. Gatica-Perez, "What did you do today? Discovering daily routines from large-scale mobile data," in *Proc. ACM Int. Conf. Multimedia (MM)*, Vancouver, BC, Canada, 2008, pp. 849–852.
[9] K. Farrahi and D. Gatica-Perez, "Learning and predicting multimodal daily life patterns from cell phones," in *ICMI-MLMI*, Cambridge, MA, 2009, pp. 277–280.
[10] N. Fay, S. Garrod, and J. Carletta, "Group discussion as interactive dialogue or serial monologue: The influence of group size," *Psychol. Sci.*, vol. 11, no. 6, pp. 487–492, 2000.
[11] M. C. Gonzalez, A. Cesar, and A. L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.
[12] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *PNAS*, vol. 101, pp. 5228–5235, 2004.
[13] G. Heinrich, parameter estimation for text analysis,' "University of Leipzig, Tech. Rep., 2008.
[14] C. A. Hidalgo and C. Rodriguez-Sickert, "The dynamics of a mobile phone network," *Phys. A*, vol. 387, pp. 3017–3024, 2008.
[15] T. Huynh, M. Fritz, and B. Schiele, "Discovery of activity patterns using topic models," in *PROC. Ubiquitous Computing (UbiComp)*, Seoul, Korea, 2008, pp. 10–19.
[16] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, Feb. 2009.
[17] L. Liao, D. Fox, and H. Kautz, "Location-based activity recognition," *Adv. Neural Inf. Process. Syst. (NIPS)*, pp. 787–794, 2006.
[18] M. Loecher and T. Jebara, "CitySense: Multiscale space time clustering of GPS points and trajectories," in *Proc. Joint Statist. Meeting*, 2009.
[19] H. Lu, W. Pan, N. Lane, T. Choudhury, and A. Campbell, "SoundSense: Scalable sound sensing for people-centric applications on mobile phones," in *Proc. 7th Int. Conf. Mobile Syst., Applicat., Services (Mobisys)*, 2009, pp. 165–178.
[20] A. Madan and A. Pentland, "VibeFones: Socially aware mobile phones," in *Proc. Int. Symp. Wearable Comput. (ISWC)*, 2006, pp. 109–112.
[21] D. O. Olguin, B. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible organizations: Technology and methodology for automatically measuring organizational behavior," *IEEE Trans. Syst., Man, Cybern.—Part B: Cybern.*, vol. 39, no. 1, pp. 43–55, Feb. 2009.
[22] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, "Learning author topic models from text corpora," *ACM Trans. Inf. Syst.*, vol. 28, no. 1, pp. 1–38, Jan. 2010.
[23] Sense Networks. [Online]. Available: http://www.sensenetworks.com
[24] P. Wang, M. C. Gonzalez, C. A. Hidalgo, and A. L. Barabasi, "Understanding the spreading patterns of mobile phone viruses," *Science*, vol. 324, no. 5930, pp. 1071–1076, May 2009.
[25] J. Zhu, A. Ahmed, and E. P. Xing, "MedLDA: Maximum margin supervised topic models for regression and classification," in *Proc. ICML*, 2009, pp. 1257–1264.

Katayoun Farrahi (M'07) received the B.Sc. degree in engineering science from the University of Toronto, Toronto, ON, Canada, and the M.Sc. degree in electrical engineering from the University of Victoria, Victoria, BC, Canada. She is currently pursuing the Ph.D. degree at the Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland, and the Idiap Research Institute, Martigny, Switzerland, under the supervision of Dr. D. Gatica-Perez.

Her general interests include machine learning, reality mining, and activity modeling.

Daniel Gatica-Perez (S'01–M'02) received the B.S. degree in electronic engineering from the University of Puebla, Puebla, Mexico, in 1993, the M.S. degree in electrical engineering from the National University of Mexico, Albuquerque, in 1996, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, in 2001.

He is now a Senior Researcher at the Idiap Research Institute, Martigny, where he directs the Social Computing Group, developing computational models to analyze human behavior from sensor data. His recent work has developed methods to analyze small groups at work in multisensor spaces, populations using cell phones in urban environments, and online communities in social media. He has published over 100 refereed papers in journals, books, and conferences in his research areas.

Dr. Gatica-Perez currently serves as an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, *Image and Vision Computing*, *Machine Vision and Applications*, and the *Journal of Ambient Intelligence and Smart Environments*, and was Guest Co-Editor of the *IEEE Computer Magazine* Special Issue on Human-Centered Computing. He received the Yang Research Award for his doctoral work.