Learning Latent Activities from Social Signals with Hierarchical Dirichlet Process

Dinh Phung, Thuong Nguyen, Sunil Gupta and S. Venkatesh

Centre for Pattern Recognition and Data Analytics, Deakin University, Australia, Waurn Ponds, VIC 3216, Australia.

Abstract

Understanding human activities is an important research topic, noticeably in assisted living and health monitoring. Beyond simple forms of activity (e.g., an RFID event of entering a building), learning latent activities that are more semantically interpretable, such as sitting at a desk, meeting with people or gathering with friends, remains a challenging problem. Supervised learning has been the typical modeling choice in the past. However, this requires labeled training data, is unable to predict never-seen-before activity and fails to adapt to the continuing growth of data over time. In this chapter, we explore the use of a Bayesian nonparametric method, in particular the Hierarchical Dirichlet Process, to infer latent activities from sensor data acquired in a pervasive setting. Our framework is unsupervised, requires no labeled data and is able to discover new activities as data grows. We present experiments on extracting movement and interaction activities from sociometric badge signals and show how to use them for detecting of sub-communities. Using the popular Reality Mining dataset, we further demonstrate the extraction of colocation activities and use them to automatically infer the structure of social subgroups.

Keywords: activity recognition, Bayesian nonparametric, hierarchical Dirichlet process, pervasive sensors, healthcare monitoring

1. Introduction

The proliferation of pervasive sensing platforms for human computing and healthcare has deeply transformed our lives. Building scalable activity-aware applications that can learn, act and behave intelligently has increasingly become an important research topic. The ability to discover meaningful and measurable activities are crucial to this endeavor. Extracting and using simple forms of activity, such as spatial whereabouts or RFID triggers of entering a building, have remained the focus of much existing work in context-aware applications (e.g., see

Email address: {dinhphung,ctng,sunil.gupta,svetha.venkatesh}@deakin.edu.au
(Dinh Phung, Thuong Nguyen, Sunil Gupta and S. Venkatesh)

Preprint submitted to Book on Plan, Activity, and Intent Recognition  September 18, 2013
a recent survey in [5, 16]). In most cases, these activities can be readily derived from the recorded raw signals: geo-location from GPS, colocation with known people from Bluetooth or the presence/absence of someone from a meeting room using RFID tags, though they might be susceptible to changes without careful data cleaning and processing when data grows [5]. Beyond these simple forms of activity, learning latent activities that are more semantically interpretable which cannot be trivially derived from the signals remains a challenging problem. We call these latent activities. For example, instead of using raw \((x, y, z)\) coordinate readings from the accelerometer sensor, we consider the movement status of the user such as walking or running; or instead of using GPS longitude/latitude readings, one wish to learn high-level daily rhythms from the GPS data [52].

Inferring these latent activities is a challenging task that has received little research attention. Most of the existing approaches have used parametric methods from machine learning and data mining. Standard supervised learning and unsupervised learning techniques are most popularly used. These include Support Vector Machines, Naive Bayes and hidden Markov models (HMM) to name a few. Typical unsupervised learning methods include Gaussian mixture models (GMM), K-means and latent Dirichlet allocation (LDA) (cf. section 2). These methods are parametric models in the sense that, once models are learned from a particular dataset, they will be fixed and unable to grow or expand as the data grows. A more severe drawback of these methods is the need to specify the number of latent activities in advance, thus limiting the model to a pre-defined set of activities. For example, one needs to specify the number of clusters for GMM, the number of states for HMM, or the number of topics for LDA. It is often difficult in practice to specify these parameters in the first place. A common strategy is to perform model selection, searching through the parameter space to pick the best performance on a held-out dataset, which can be cumbersome, computationally inefficient and sensitive to the held-out dataset. These parametric methods are potentially unsuitable for ubiquitous applications, such as in pervasive health, as the data is typically growing over time and has “no clear beginning and ending” [1].

In this chapter, we address these problems by proposing the use of Bayesian nonparametric methods (BNP), a recent data modeling paradigm in machine learning, to infer latent activities from sensor data acquired in a pervasive setting with an emphasis on social signals. These Bayesian models specify a nonparametric prior on the data generative process; as such, they allow the number of latent activities to be learned automatically and grow with the data, hence overcome the fundamental difficulty of model selection encountered in those aforementioned parametric models. In particular, we employ the hierarchical Dirichlet process (HDP) [62], a recent BNP model designed for modeling grouped data such as a collection of documents, images or signals represented in appropriate form. Similar to other Bayesian nonparametric models, HDP automatically infers the number of latent topics – equivalent to our latent activities – from data and assign each data point to a topic; hence HDP simultaneously discovers latent activities and assigns a mixture proportion to each document (or a group of signals), which can be used as a latent feature representation for
users’ activities for additional tasks such as classification or clustering.

We demonstrate the framework on the collection of social signals, a new type of data sensing paradigm and network science pioneered at MIT Media Lab [50]. We experiment with two datasets. The first was collected in our lab using sociometric badges provided by Sociometric Solutions\(^1\). The second dataset is the well-known Reality Mining dataset [20]. For the sociometric data, we report two experiments: discovering latent movement activities derived from accelerometer data and colocation activities from Bluetooth data. To demonstrate the feasibility of the framework in a larger setting, we experiment with the Reality Mining dataset and discover the interaction activities from this data. In all cases, to demonstrate the strength of Bayesian nonparametric frameworks, we also compare our results with activities discovered by latent Dirichlet allocation [7] — a widely acknowledged state-of-the-art parametric model for topic modeling. We derive the similarity between subjects in our study using latent activities discovered, then feed the similarity to a clustering algorithm and compare the performance between the two approaches. We quantitatively judge the clustering performance on four metrics commonly used in information retrieval and data mining: F-measure, Rand index, purity and normalized mutual information (NMI). Without the need of searching over the parameter space, HDP consistently delivers result comparable with the best performance of LDA, up to 0.9 in F-measure and NMI in some cases.

2. Related Work

We shall briefly describe a typical activity recognition system before surveying work on sensor-based activity recognition using wearable sensors. Background related to Bayesian nonparametric models is deferred to the next section.

Activity recognition systems

A typical activity recognition system contains three components: a sensor infrastructure, a feature extraction component, and a learning module for recognizing known activities to be used for future prediction.

The sensor infrastructure uses relevant sensors to collect data from the environment and/or people. A wide range of sensor types have been used in the past decade. Early work in activity recognition has mainly focused on using static cameras (e.g. [11, 18, 19]) or circuits embedded indoors [64, 65]. However, these types of sensors are limited to some fixed location and only provide the ability to classify activities that have occurred previously. Often they cannot be used to predict future activities, which is more important, especially in health care assistant systems.

Much of the recent work has shifted focus to wearable sensors due to their convenience and ease of use. Typical sensing types include GPS [40], accelerometers [24, 30, 36, 38, 46, 57], gyroscopes [24, 38], galvanic skin response (GSR)

\(^1\)http://www.sociometricsolutions.com/
sensor, electrocardiogram (ECG) sensor \([33, 38]\) to detect body movement and physical states. These sensors have been integrated into pervasive commercial wearable devices such as Sociometric Badge\(^2\) \([47]\) or SHIMMER\(^3\) \([10]\). These types of sensors and signals are becoming more and more prevalent, creating new opportunities and challenges in activity recognition research.

The purpose of the *feature extraction* component is to extract the most relevant and informative features from the raw signals. For example, in vision-based recognition, popular features include SIFT descriptors, silhouettes, contours, edges, pose estimates, velocities and optical flow. For temporal signals such as those collected from accelerometers or gyroscopes, one might use basic descriptive statistics derived from each window such as mean, variance, standard deviation, energy, entropy, FFT and wavelet coefficients \([2, 57, 59]\). Those features might further be processed to remove noise or reduce the dimensionality of the data using machine learning techniques such as Principal Component Analysis (PCA) \([31]\) or Kernel PCA \([66, 68]\), to provide a better description of the data.

The third, and perhaps the most important component of an activity recognition system is the *learning module*. This module learns predefined activities from training data with labels to be used for future activity prediction. Activity labels are often manually annotated during the training phase and supplied to a supervised learning method of choice. Once trained, the system can start to predict the activity label based on acquired signals. However, manually annotating activity labels is a time-consuming process. In addition, the activity patterns may grow and change over time, and this traditional method of supervised training might fail to address real-world problem. Our work presents an effort to address this problem. A relevant chapter in this book by Rashidi \([56]\) also recognizes this challenge and proposes a sequential data mining framework for human activity discovery. However, Rashidi’s work focuses on discovering dynamic activities from raw signals collected from a data stream, whereas by using Bayesian nonparametric models, our work aims to discover higher-order latent activities such as *sitting*, *walking*, *people gathering*, referred to as *situations* or *goals* in Rashidi’s chapter.

### Pervasive sensor-based activity recognition

As mentioned earlier, much of recent work has shifted focus to sensor-based approaches due to the rapid growth of wearable devices. In particular, accelerometer signals have been exploited in pervasive computing research (e.g., \([36, 46, 57, 9]\)). Ravi et al. \([57]\) experimented with a single 3-axis accelerometer worn near the pelvic region and attempt to classify a set of eight different activities. Olguin and Pentland \([46]\) defined and classified eight different activities using data generated by 3 accelerometers placed on right wrist, left hip and chest using the HMM. Kwapisz et al. \([36]\) used accelerometer data from

---

\(^2\)http://www.sociometricsolutions.com/

\(^3\)http://www.shimmer-research.com/
cell phones of 29 users to recognize various activities and gestures. Unlike our work in this chapter, a common theme of these approaches is that they employ supervised learning which requires labeled data.

Another thread of work relevant to ours uses wearable signals to assess energy expenditure (e.g., [24]). The atomic activity patterns derived in our work might have implications in this line of research. In contrast to [24], our results on body movement pattern extraction do not require any supervision. In addition to accelerometer data, signals from multi-sensor model have also been investigated. Choudhury et al. [12] constructed a sensing platform that includes six types of sensors for activity recognition; again with a supervised learning setting they report good results when accelerometer data is combined with microphone data. Notably, Lorincz et al. [38] introduced a platform, called Mercury, to integrate a wearable sensor network into a unified system to collect and transfer data in real time or near real time. These types of platform are complementary to the methods proposed in this chapter for extracting richer contexts and patterns.

**Pervasive sensors for healthcare**

Pervasive sensors have been popularly used in much of the recent research in healthcare. This research can be classified into three categories: anomaly detection in activity, physical disease detection and mental disease detection. Anomaly in human activity might include falls by aged or disabled patients, changes in muscle/motion function, and detection of those anomalies can be performed using accelerometers [14, 60, 63], wearable cameras [60], barometric pressure sensors [63], or passive infrared motion sensors [67]. Real time management of physical diseases such as cardiac or respiratory disease can improve the diagnosis and treatment. Much of the recent research utilizes wearable electrocardiogram (ECG) sensors [4, 13] or electromagnetic generators [48]. However, in our opinion, the most difficult and challenging research problem related to healthcare is the mental problem detection. Mental problems might include emotion, mood, stress, depression or shock. Much recent work has focused on using physiological signals such as ECG [26] or galvanic skin response (GSR) [39, 55]. Some other researchers also have examined the utility of audio signals [39] to detect stress and this problem remains challenging today due to the lack of psychoacoustic understanding of the signals.

**Human dynamics and social interaction**

Understanding human dynamics and social interaction from a network science perspective presents another line of related work. Previous research on discovery of human dynamics and social interaction can be broadly classified into two groups based on the signal used: base-station signal and pairwise signal. In the first group, the signal is collected by wearable devices to extract their relative position to some base stations with known positions. This approach might be used to perform indoor positioning [3, 28, 34]. The work of [3] employs this approach and introduces the Context Management Frame infrastructure, in which the Bluetooth fingerprint is constructed by received signal strength
indication (RSSI) and response rate (RR) to base-stations. This requires some training data with the location-labeled fingerprint in order to extract the location of new fingerprint data. Kondragunta [34] built an infrastructure system with some fixed base stations to detect wearable Bluetooth devices. They used the particle filtering method to keep track of people and detect the location contexts. However, these approaches share the same drawback as they require some infrastructure to be installed in advance to fixed locations, and cannot be easily extended to new locations. Our early work [51, 53, 52] attempted to model high-level routine activities from GPS and WiFi signals, but remains limited to a parametric setting.

Some recent work has attempted to extract high-level activities and perform latent context inference (e.g., [30], [33]). The authors in [30] applied latent Dirichlet allocation (LDA) to extract daily routine. They used a supervised learning step to recognize activity patterns from data generated by 2 accelerometers and then applied a topic model on discovered activity patterns to recognize daily routines. Unsupervised clustering algorithm (K-means) was used to cluster sensor data into $K$ clusters and use these clusters as the vocabulary for topic modeling. Kim et al. [33] attempted to extract high level descriptions of physical activities from ECG and accelerometer data using latent Dirichlet allocation (LDA). These approaches are parametric in nature, requiring many key parameters to be specified in advance, such as the number of topics or the size of vocabulary. In fact, methods presented in our work can be seen as a natural progression to the work of [30] and [33] in which they overcome those aforementioned drawbacks.

The second related line of work is detecting colocation context based on pairwise Bluetooth signals. In this approach, each user wears a device that can detect other surrounding devices within a short distance. From this type of signal, a pairwise colocation network can be constructed and we can use a graph-based algorithm to extract sub-communities from this network. A suitable method is $k$-clique proposed by [15, 49]. This method attempts to extract $k$-cliques – sub graphs of $k$ fully connected nodes – from the network. The $k$-clique in this approach might be seen as a colocation context in our approach. Kumpula et. al. [35] proposed a sequential method for this approach. However, to be able to use these methods, we have to generate a pairwise colocation network and the corresponding weights. Instead of a graph-based approach, [8, 21, 42, 44] use the signal directly to detect colocation groups. Nicolai & Kenn [44] collected proximity data via Bluetooth of a mobile phone during the Ubicomp conference in 2005. They count the number of familiar and unfamiliar people and compute the dynamic of the familiar and unfamiliar group. These dynamic metrics reflect the events during the conference. They have not proposed a method to detect typical colocation groups. Eagle et al. [21] use the Reality Mining dataset [20] to analyze users’ daily behaviors using a factor analysis technique, where the behavior data of each user in a day can be approximated by weighted sum of some basic behavioral factors named eigenbehaviors. They employ this method which is a parametric model with different number of factors and run in a batch setting. Dong et al. [17] use the Markov jump
process to infer the relationships between survey questions and sensor data on a dataset collected by mobile phones. Their work includes two aspects: predicting relations (friendship) from sensor data and predicting sensor data from the survey questions.

3. Bayesian Nonparametric Approach to Inferring Latent Activities

We first briefly discuss probabilistic mixture models to motivate the need of the Dirichlet process to specify infinite mixture modeling. Dirichlet process mixture model (DPM) is then described followed by the hierarchical Dirichlet process.

3.1. Bayesian mixture modeling with Dirichlet processes

Mixture modeling is a probabilistic approach to model the existence of latent sub-populations in the data. A mixture component \( k \) is represented by a cluster-specific\(^4\) probability density \( f_k(\cdot | \phi_k) \) where \( \phi_k \) parametrizes the cluster distribution. For a data point \( x \), its sub-population is assumed to be unknown, hence the generative likelihood is modeled as a mixture from \( K \) sub-populations:

\[
p(x | \pi, \phi_{1:K}) = \sum_{k=1}^{K} \pi_k f(x | \phi_k)
\]

where \( \pi_k \) is the probability that \( x \) belongs to the \( k \)-th sub-population and \( \sum_{k=1}^{K} \pi_k = 1 \). This is the parametric and frequentist approach to mixture modeling. The EM algorithm is typically employed to estimate the parameters \( \pi \) and \( \phi_k \) from the data.

Under a Bayesian setting the parameters \( \pi \) and \( \phi_k \) are further endowed with prior distributions. Typically a symmetric Dirichlet distribution \( \text{Dir}(\gamma) \) is used as the prior of \( \pi \), and the prior distributions for \( \phi_k \) are model-specific depending on the form of the likelihood distribution \( F \). Conjugate priors are preferable; for example, if \( F \) is a Bernoulli distribution, then conjugate prior distribution \( H \) is a Beta distribution. Given the prior specification and defining \( f \) as the density function for \( F \), Bayesian mixture models specify the generative likelihood for \( x \) as:

\[
p(x | \gamma, H) = \int \int_{\phi_{1:K}} \sum_{k=1}^{K} \pi_k f(x | \phi_k) dP(\pi)dP(\phi_{1:K})
\]

Under this formalism, inference is required to derive the posterior distribution for \( \pi \) and \( \phi_k \) which is often intractable. Markov Chain Monte Carlo methods, such as Gibbs sampling, are common approaches for the task. A latent indicator

\(^4\)We also use the term ‘probability density’ to imply a ‘probability mass function’ when the data domain is discrete.
variable $z_i$ is introduced for each data point $x_i$ to specify its mixture component. Conditional on this latent variable, the distribution for $x_i$ is simplified to:

$$ p(x_i \mid z_i = k; \phi_1, K, \pi) = \pi_k f(x_i \mid \phi_k) \tag{3} $$

where $\Pr(z_i = k) = \pi_k$. Full Gibbs sampling for posterior inference becomes straightforward by iteratively sampling the conditional distributions among the latent variable $\pi, z_i$ (s) and $\phi_k$ (s), i.e.,

$$ p(z_i \mid z_{-i}, x_{1:n}, \pi, \phi_1, K) \propto p(x_i \mid z_i) = f(x_i \mid \phi_{z_i}) \tag{4} $$

$$ p(\pi \mid z_{1:n}, x_{1:n}, \phi_1, K) \propto p(z_{1:n} \mid \pi) p(\pi) \tag{5} $$

$$ p(\phi_k \mid z_{1:n}, x_{1:n}, \pi, \phi_{-k}) \propto \prod_{x \in \{x ; z_i = k, i = 1, \ldots, n\}} p(x_i \mid \phi_k) p(\phi_k) \tag{6} $$

It is not hard to recognize that the posterior for $\pi$ is again a Dirichlet, and with a conjugate prior, the posterior for $\phi_k$ also remains in the same form, hence they are straightforward to sample. A collapsed Gibbs inference scheme can also be developed to improve the variance of the estimators by integrating out $\pi$ and $\phi_k$ (s), leaving out the only following conditional to sample from:

$$ p(z_i = k \mid z_{-i}, x_{1:n}, \Psi) \propto (\alpha + n^{-1}_k) \int_{\phi_k} p(x_i \mid \phi_k) p(\phi_k \mid \{x_j : j \neq i, z_j = k\}, \Psi) d\phi_k \tag{7} $$

where $n^{-1}_k = \sum_{j=1,j\neq i}^n \mathbb{1}\{z_j = k\}$ is the number of assignments to cluster $k$, excluding position $i$ and $\alpha$ is the hyperparameter for $\pi$, assumed to be a symmetric Dirichlet distribution. The second term involves an integration which can easily be recognized as the predictive likelihood under a posterior distribution for $\phi_k$. For conjugate prior, this expression can be analytically evaluated. Several results can readily be found in many standard Bayesian text books such as [27].

A key theoretical limitation in the parametric Bayesian mixture model described so far is the assumption that the number of sub-populations in the data is known. In the realm of activity recognition, given the observed signals $x_i$ (s), each sub-population can be interpreted as one class of activity and it is desirable to infer the number of classes automatically. Model selection methods such as cross-validation, Akaike information criterion (AIC) or Bayesian information criterion (BIC) can be used to estimate $K$. However, these methods are still sensitive to held-out data. Moreover, it is expensive and difficult to re-select the model when new data arrives.

Recent advances in Bayesian nonparametric modeling (BNP) provides a principled alternative to overcome these problems by introducing a nonparametric

---

5Here we assume to observe $n$ data points $x_{1:n}$ and we write $z_{-i}$ to imply the set of all $z$ (s) except $z_i$, i.e., $z_{-i} = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$; a similar convention is applied for $\phi_{-k}$.
prior distribution on the parameters, hence allowing a countably infinite number of sub-populations, leading to an \textit{infinite mixture} formulation for data likelihood (cf. Eq 1)

\begin{equation}
p(x \mid \cdot) = \sum_{k=1}^{\infty} \pi_k f(x \mid \phi_k)
\end{equation}

It is then natural to ask what would be a suitable prior distribution for infinite-dimensional objects that arise in the above equation. The answer to this is the Dirichlet process [23]. One way to motivate the arrival of the Bayesian nonparametric setting is to reconsider the mixture likelihood in Eq (1). As usual, let $(\pi_1, \ldots, \pi_K) \sim \text{Dir}(\gamma), \phi_k \overset{iid}{\sim} H, k = 1, \ldots, K$, we construct a new discrete measure:

\begin{equation}
G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}
\end{equation}

where $\delta_{\phi_k}$ denotes the atomic probability measure placing its mass at $\phi_k$. The generative likelihood for data $x$ can be equivalently expressed as: $x \mid \phi \sim F(x \mid \phi)$ where $\phi \mid G \sim G$. To see why, we express the conditional distribution for $x$ given $G$ as

\begin{align}
p(x \mid G) &= \int_{\phi} f(x \mid \phi) dG(\phi) = \int_{\phi} f(x \mid \phi) \sum_{k=1}^{K} \pi_k \delta_{\phi_k} d\phi \\
&= \sum_{k=1}^{K} \pi_k \int_{\phi} f(x \mid \phi) \delta_{\phi_k} d\phi = \sum_{k=1}^{K} \pi_k f(x \mid \phi_k)
\end{align}

which identically recovers the likelihood form in Eq (1). Hence, one sensible approach to move from a finite mixture representation in Eq (1) to an infinite mixture in Eq (8) is to define a discrete probability measure constructed from an infinite collection of atoms $\phi_k (s)$ sampled iid from the base measure $H$ and the weight $\pi_k (s)$:

\begin{equation}
G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}
\end{equation}

Since both $\phi_k (s)$ and $\pi_k (s)$ are random, $G$ is also a random probability measure. It turns out that random discrete probability measures constructed by this way follow a Dirichlet process. Briefly, a Dirichlet process DP $(\gamma, H)$ is a distribution over random probability measure $G$ on the parameter space $\Omega$ and is specified by two parameters: $\gamma > 0$ is the \textit{concentration} parameter and $H$ is base measure [23]. The terms ‘Dirichlet’ and ‘base measure’ come from the fact that for any finite partition of the measurable space $\Omega$, the random vector obtained by applying $G$ on this partition will distribute according to a Dirichlet distribution with parameters obtained from $\gamma H$. Alternatively, the nonparametric object
G in Eq (12) can be viewed as a limiting form of the parametric G in Eq (9) when $K \to \infty$ and weights $\pi_k(s)$ are drawn from a symmetric Dirichlet $\pi \sim \text{Dir}(\gamma_1, \ldots, \gamma_K)$ [62]. Using G as a nonparametric prior distribution, the data generative process for infinite mixture models can be summarized as follows:

\[ G \sim \text{DP}(\gamma, H) \]  
\[ \theta_i \sim G, \quad \text{for } i = 1, \ldots, n \]  
\[ x_i \sim F(\cdot | \theta_i) \]

The representation for G in Eq (12) is known as the stick-breaking representation, taking the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ where $\phi_k \sim H, k = 1, \ldots, \infty$ and $\pi = (\pi_k)_{k=1}^{\infty}$ are the weights constructed through a ‘stick-breaking’ process $\pi_k = v_k \prod_{s<k} (1 - v_s)$ with $v_k \sim \text{Beta}(1, \gamma), k = 1, \ldots, \infty$ [58]. It can be shown that $\sum_{k=1}^{\infty} \pi_k = 1$ with probability one, and we denote this process as $(\pi_1, \pi_2, \ldots) \sim \text{GEM}(\gamma)$. A recent book [29] provides an excellent account on theory and applications of DP.

Inference in the DPM can be carried out under a Gibbs sampling scheme using the Polya urn characterization of Dirichlet process [6], otherwise also known as the Chinese restaurant process [54]. Again, let $z_i$ be the cluster indicator variable for $x_i$, the collapsed Gibbs inference reduces to [43]:

\[ p(z_i = k | z_{-i}, x_{1:N}, \gamma) \propto \begin{cases} \frac{(n_{k}^{-1})}{f_k^{-x_i}}(x_i) & \text{if } k \text{ previously used} \\ \gamma f_k^{-x_i}(x_i) & \text{if } k \text{ takes on a new value} \end{cases} \]  

where $n_{k}^{-1}$ denotes the number of times the k cluster has been used excluding position $i$ and $f_k^{-x_i}(x_i) = \int_{\phi_k} p(x_i | \phi_k) p(\phi_k | \{x_j : z_j = k, j \neq i\}) d\phi_k$ is the predictive likelihood of observing $x_i$ under the k-th mixture, which can be evaluated analytically thanks to the conjugacy between $H$ and $F$. The expression in Equation (16) illustrates the clustering property induced by DPM: a future data observation is more likely to return to an existing cluster with a probability proportional to its popularity $n_{k}^{-1}$, but it is also flexible enough to pick on a new value if needed as data grows beyond the complexity that current model can explain. Furthermore, the number of clusters grow at $O(n \log \gamma)$ under the Dirichlet process prior.

### 3.2 Hierarchical Dirichlet Process

Activity recognition problems typically involve multiple sources of information either from an egocentric or device-centric point of view. In the former, the individual user acquires user-specific datasets whereas in the later different sensors collect their own observations. Nonetheless, these data sources are often correlated by an underlying and hidden process that dictates how the data is observed. It is desirable to build a hierarchical model to learn activities from these multiple sources jointly by leveraging mutual statistical strength across datasets. This is known as the ‘shrinkage’ effect in Bayesian analysis.
The Dirichlet process mixture model described in the previous section is suitable for analyzing a single data group, such as data that comes from an individual user or a single device. Moreover, the Dirichlet process can also be utilized as a nonparametric prior for modeling grouped data. Under this setting, each group is modeled as a DPM and these models are ‘linked’ together to reflect the dependency among them.

Figure 1: Left: stick-breaking graphical model representation for HDP in plate notation. Right: expanded graphical model representation over groups. Each box represents one group of data (e.g., document) whose observed data points are shaded. Unshaded nodes are latent variables. Each observed data point $x_{ji}$ is assigned to a latent mixture indicator variable $z_{ji}$; $\gamma$ and $\alpha$ are the concentration parameters and $H$ is the base measure. Each group of data possesses a separate mixture proportion vector $\pi_j$ of latent topics $\phi_k$.

One particularly attractive formalism is the hierarchical Dirichlet process [61, 62] which posits the dependency among the group-level DPM by another Dirichlet process. Due to the discreteness property of the DP, mixture components are naturally shared across groups. Specifically, let $J$ be the number of groups and $\{x_{j1}, \ldots, x_{jN_j}\}$ be $N_j$ observations associated with the group $j$ which are assumed to be exchangeable within the group. Under the HDP framework, each group $j$ is endowed with a random group-specific mixture distribution $G_j$ which is hierarchically ‘linked’ by another DP with a measure measure $G_0$, who itself is a draw from another DP:

$$G_j \mid \alpha, G_0 \sim \text{DP} (\alpha, G_0), j = 1, \ldots, J$$  (17)

$$G_0 \mid \gamma, H \sim \text{DP} (\gamma, H)$$  (18)

The stick-breaking construction for HDP makes it clear how mixture compo-
nents are shared across groups:

\[ G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \text{ where } (\beta_1, \beta_2, \ldots) \sim \text{GEM} (\gamma), \quad \phi_k \sim \text{iid } H \quad (19) \]

\[ G_j = \sum_{k=1}^{\infty} \pi'_{jk} \delta_{\psi_k} \text{ where } (\pi'_{j1}, \pi'_{j2}, \ldots) \sim \text{GEM} (\alpha), \quad \psi_k \sim \text{G}_0 \quad (20) \]

Since \( G_0 \) is discrete, taking the support on \( \{\phi_1, \phi_2, \ldots\} \), \( G_j \) can equivalently be expressed \[62\] as:

\[ G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \text{ where } \pi_j | \alpha, \beta \sim \text{DP} (\alpha, \beta) \quad (21) \]

Recall that \( x_{ji} \) and \( x'_{ji} \) belong to the same mixture component if they are generated from the same component parameter, i.e., \( \theta_{ji} = \theta'_{ji} \) (cf. Figure 1).

Equations (19) and (21) immediately reveal that \( \theta_{ji}(s) \) share the same discrete support \( \{\phi_1, \phi_2, \ldots\} \), hence mixture components are shared across groups.

Integrating out the random measures \( G_j(s) \) and \( G_0 \), the posterior for \( \theta_{ji}(s) \) has been shown to follow a Chinese Restaurant Franchise process which can be utilized to develop inference algorithm based on Gibbs sampling and we refer to \[62\] for details. Alternatively, an auxiliary variable collapsed Gibbs sampling can be developed based on the stick-breaking representation by integrating out group-specific stick-breaking weights \( \pi_{jk}(s) \) and the atomic atoms \( \phi_k(s) \) but explicitly sampling the stick weights \( \beta_k(s) \) at the higher level as described in \[62\].

Assume \( L \) is the number of clusters have been created, but only \( K < L \) components are active, which means the cluster indicator is limited within \( K \) clusters \( z_{ji} \in \{1, \ldots, K\} \). There are \( (L-K) \) inactive components, which we shall cluster them together and represented by \( k_{\text{new}} \). Therefore, using the additive property of Dirichlet distribution, \( (\beta_1, \ldots, \beta_K, \beta_{K+1}, \ldots, \beta_L) \sim \text{Dir} (\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}) \) can now be written as \( (\beta_1, \ldots, \beta_K, \beta_{\text{new}}) \sim \text{Dir} (\gamma_1, \ldots, \gamma_K, \gamma_{\text{new}}) \) where \( \gamma_k = \frac{\alpha}{K} \) for \( k = 1, \ldots, K \) and \( \gamma_{\text{new}} = \frac{(L-K)\alpha}{K} \). Integrating out \( \pi_{1,j} \) and \( \phi_k(s) \) our Gibbs sampling state space is \( \{z, \beta_{1:K}, \beta_{\text{new}}\} \). To sample \( z_{ji} \), since there are only \( K \) current active components, we can work out the conditional distribution for \( z \) by discarding inactive components \( \beta_{\text{new}} \) and integrating out \( \pi_j \) within each group \( j \) where we note that \( \pi_j \sim \text{Dir} (\alpha \beta_1, \ldots, \alpha \beta_K) \), hence we apply the marginal likelihood result for the Multinomial-Dirichlet conjugacy:

\[ p(z | \beta_{1:L}) = p(z | \beta_{1:K}) = \prod_{j=1}^{K} p(z_j | \beta_{1:K}) \quad (22) \]

\[ = \prod_{j=1}^{K} \int_{\pi_j} p(z_j | \pi_j) p(\pi_j | \beta_{1:K}) d\pi_j \quad (23) \]

\[ = \prod_{j=1}^{K} \frac{\Gamma(\alpha + N_j) \prod_{k=1}^{K} \Gamma(\alpha \beta_k + n_{jk})}{\Gamma(\alpha) \prod_{k=1}^{K} \Gamma(\alpha \beta_k)} \quad (24) \]

12
With some further simple manipulations, we get: \( p(z_{ji} = k | z^{-ji}, \beta_{1:K}) \propto (\alpha \beta_k + n_k^{-ji}) \). Therefore, together with the data \( x_{ji} \), the Gibbs sampling equation for \( z_{ji} \) can be established:

\[
p(z_{ji} = k | z^{-ji}, \beta_{1:K}, x) \propto p(z_{ji} = k | z^{-ji}, \beta_{1:K}) \times \int p(x_{ji} | \phi_k) p(\phi_k | \{x_{j'i'} : z_{j'i'} = k, (j', i') \neq (j, i)\}) d\phi_k
\]

(25)

Again the last term is a form of predictive likelihood for \( x_{ji} \) under a standard Bayesian setting. It is model-specific but can be evaluated analytically when \( H \) is conjugate to \( F \). Note that when resampling \( z_{ji} \) we have to consider the case it may take on a new value \( k_{\text{new}} \). If \( z_{ji} = k_{\text{new}} \) introduces a new active cluster, we then need to update as follows: \( K \leftarrow K + 1, z_{ji} = K, \phi_{K+1} = \phi_{\text{new}}, \beta_K \leftarrow \epsilon \beta_{\text{new}}, \beta_{\text{new}} \leftarrow (1 - \epsilon) \beta_{\text{new}} \) where \( \epsilon \sim \text{Beta} (1, \gamma) \) where the splitting step for \( \beta_K \) and \( \beta_{\text{new}} \) is justified by the splitting property of the Dirichlet distribution [62].

To sample \( \{\beta_{1:K}, \beta_{\text{new}}\} \), we express their conditional distribution, which turns out to be the ratio of two Gamma functions, hence, as shown in [62], we sample them together with an auxiliary variable \( m_{jk} \) where note that conditional on \( m, \beta \) follows a Dirichlet distribution:

\[
q(m_{jk} = m | z, m^{-jk}, \beta) \propto s(n_{jk}, m) (\alpha \beta_k)^m
\]

(27)

\[
q(\beta_{1:K}, \beta_{\text{new}} | z, m) \propto \beta_{\text{new}}^{\gamma - 1} \prod_{k=1}^K \beta_k^{m_{jk} - 1} \]

(28)

The hyper-parameters \( \alpha \) and \( \gamma \) can also be resampled by further endowing them with Gamma distributions and following the methods described in [22]. In this work, these hyper-parameters are always resampled together with the main parameters. Lastly, assuming that the size of each group is roughly the same at \( O(N) \), as shown in [61], the number of mixture components grows at \( O\left(\gamma \log \frac{N}{\gamma} + \gamma \log J + \gamma \log \log \frac{N}{\alpha}\right) \) which is doubly logarithmically in \( N \) and logarithmically in \( J \).

4. Experiments

Two datasets are used: the first collected in our lab using sociometric badges provided by Sociometric Solutions\(^6\) and the second is the well-known public Reality Mining data [20].

\(^6\)http://www.sociometricsolutions.com/
4.1. Sociometric Data

Sociometric badges were invented by MIT Media Lab [47] and have been used in various applications. Figure 2 shows the front of a badge which shows the infrared sensor, the top-side microphone\(^7\) and the programming mode button. These badges record a rich set of data, including: anonymized speech features (e.g., volume, tone of voice, speaking time); body movement features derived from accelerometer sensors (e.g., energy and consistency); information of people nearby and face-to-face interaction with those who are also wearing a sociometric badge, the presence and signal strength of Bluetooth-enabled devices and approximate location information. We briefly explain these features below.

![Figure 2: The front appearance of the sociometric badge and its software user interface.](image)

**Body movement**: each badge has a 3-axis accelerometer, set at a sampling rate of every 50ms. For our data collection, a window frame of 10s was used and the data was averaged within each window to construct one data point. We record the coordinate \((x, y, z)\) and two derived features: *energy* and *consistency*. Energy is simply the mean of 3-axis signals \(\sqrt{x^2 + y^2 + z^2}\) and the consistency attempts to capture the stability or consistency of the body movement computed as \(1 - \alpha\) where \(\alpha\) is the standard deviation of the \(\sqrt{x^2 + y^2 + z^2}\) values. These two features have been experimented previously with sociometric badges and reported to be good features [45].

**Proximity and people nearby**: the badge has the capacity to record the opportunistic devices with Bluetooth-enabled. It can detect the devices’ MAC address as well as the received signal strength indicator (RSSI).

We ordered 20 sociometric badges from Sociometric Solutions\(^8\), collecting data for 3 weeks on every working day, 5 days a week. However, due to some inconsistency in the way members of the lab were carrying and collecting data...

---

\(^7\)The purpose of the top-side microphone is to detect the speech by the user wearing the badge. Full descriptions of these badges can be obtained from the company website at: http://www.sociometricsolutions.com/

---

\(^8\)Sociometric Solutions
at the initial phase, we cleaned up and reported data from only 11 subjects in this chapter. We focus on two types of data in this chapter: accelerometer data and Bluetooth data. We refer to the accelerometer data as *activity data* and the Bluetooth data as *colocation data*.

### 4.1.1. Activity data

The sociometric badges record activity data from a 3-axis accelerometer. We used the *consistency* measure [45] as the activity data representing the body movement. A consistency value closer to 1 indicates less movement and closer to 0 if more movement was detected. We aim to extract basis latent activities from these signals which might be used to build healthcare applications such as monitoring users with obesity or stress. Using Bayesian nonparametric approach, we treat each data point as a random draw from a univariate Gaussian with unknown mean ($\mu$) and unknown precision ($\tau$):

$$p(x | \mu, \sigma^2) = \mathcal{N}(x | \mu, \tau^{-1}) = \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x - \mu)^2\right\}$$  \hspace{0.5cm} (29)

Each Gaussian, parametrized by ($\mu, \tau$), *represents one latent activity*. We further specify the conjugate prior distribution for ($\mu, \tau$), which is a product of Gaussian-Gamma:

$$p(\tau | \alpha_0, \beta_0) = \text{Gamma} (\tau | \alpha_0, \beta_0)$$  \hspace{0.5cm} (30)

$$p(\mu | \mu_0, s_0, \tau) = \mathcal{N}(\mu | \mu_0, (s_0 \tau)^{-1})$$  \hspace{0.5cm} (31)

where $\mu_0, \alpha_0, \beta_0$ and $s_0$ are hyper-parameters for the prior distribution in a standard Bayesian setting. These hyper-parameters express the prior belief on the distribution of the parameter space and only need to be specified roughly [27]. A widely used approach is to carry out an empirical Bayes step using observed data to infer these parameters before running Gibbs. We adapt this strategy in our chapter.

For the activity data, values of a user on a day are grouped into a document, where each data point is the consistency value in a 10-second interval during the collection time. We run Gibbs sampling for 250 iterations with a burn-in period of 50 samples. The concentration parameters $\gamma$ and $\alpha$ are also resampled according to a Gamma distribution as described in [62]. Figure 3a plots the value $\alpha$ for each Gibbs iteration, settling at around 1.2. Figure 3b shows the posterior distribution over the number of latent activities $K$ whose mode is 4; and Figure 3c plots the estimated value of $K$ together with its mode as we run the Gibbs sampler. In this case, the model has inferred $K = 4$ latent activity activity, each is represented by a univariate Gaussian distribution. These four activity are shown on the left-hand side of Figure 4, which can be seen to be separated and capturing the activity level well. The right-hand side of Figure 4 illustrates an example of consistency signal from subject 3 with his annotated activity.
Figure 3: HDP results run for activity data acquired from sociometric badges from 11 subjects. a) shows the concentration parameter $\alpha$ being sampled at each Gibbs iteration, which tends to settle down around 1.2, b) the empirical posterior distribution for the number of latent activities inferred, and c) the detailed value of $K$ and its mode tracked from the posterior distribution as we run Gibbs.

Figure 4: LHS: our model has learned four latent activities, each of which is represented by a univariate Gaussian: $\mu_1 = 0.992, \sigma^2_1 = 0.0015; \mu_2 = 0.837, \sigma^2_2 = 0.0049; \mu_3 = 0.693, \sigma^2_3 = 0.0050; \mu_4 = 0.493, \sigma^2_4 = 0.0051$. y-axis represents the consistency value derived from the $(x, y, z)$ accelerometer sensor described in section 4.1. A learned latent activity implies a moving status of the subject. For example, activity 1 has a mean close to 1, hence represents the state of non-movement, e.g., sitting. RHS: an example of consistency signal from subject 3 with annotated activities. The decoded sequence of states is also plotted underneath. This decoded sequence was computed using the MAP estimation from the Gibbs samples, each state assumes a value in the set $\{1, 2, 3, 4\}$ corresponding to 4 latent activities discovered by HDP.
Figure 5: Left: examples of raw Bluetooth readings showing time of the event, user ID and Bluetooth signal strengths; Middle: an example of Bluetooth data segment acquired for user 2 using the binary representation described in the text; Right: five atomic latent colocation activities discovered by HDP, each of which is a multinomial distribution over 11 subjects (the darker, the higher probability, e.g., colocation activity 4 shows the colocation of subjects 2, 4, 7 and 11 as can also visually be seen in data from user 2). Room number shows the actual room each subject is sitting (e.g., user 11 sits in room 236).

4.1.2. Colocation data

The sociometric badge regularly scans for surrounding Bluetooth devices including other badges. Therefore, we can derive the colocation information of the badge user. We used a window of 30 seconds to construct data in the form of count vectors. In each time interval, a data point is constructed for each user whose badge detects other badges. There are $V = 11$ valid subjects in our sociometric dataset and thus each data point is a 11-dimensional vector whose element represents the corresponding presence status of the subject and the owner of the data point is always present. Data from each day of a user is grouped into one document. All documents from all subjects are collected together to form a corpus to be used with the HDP (thus each subject may correspond to multiple documents). We model the observation probability as a multinomial distribution (the distribution $F$ in our description in Section 3.1) whose conjugate prior is the Dirichlet (the base measure $H$). Thus, a latent colocation activity is described by a discrete distribution $\phi = (\phi_1, \ldots, \phi_V)$ over the set $\{1, 2, \ldots, V = 11\}$ satisfying $\sum_{v=1}^{V} \phi_v = 1$ and can be interpreted as follows: the higher probability of an entry represents the higher chance of its colocation with others. For example, $\phi_1 = \phi_2 = 0.5$ and $\phi_v = 0$ everywhere else indicates a strong colocation activity of subject 1 and 2.

Figure 6 plots various statistics from our Gibbs inference which shows that our model has inferred 5 latent colocation activities from the data as shown in Figure 5. These activities are interpretable by inspecting against the ground truth from the lab. For example, the first activity represents the colocation of PhD students because they share the same room; the second captures only two subjects which happen to be two research fellows sharing the same room.

To further quantify the results from our model and demonstrate its strength over existing parametric models, we compare our model against latent Dirichlet
allocation [7] which has popularly been used as the state-of-the-art method for activity profiling and modeling in pervasive settings [30, 51, 53]. For each subject, we derive an average mixture proportion of latent activities (topics) from $\pi_j$ (s) vectors learned from the data for that subject. We then use Jensen-Shannon divergence to compute the distance between each pair of subjects. The Jensen-Shannon divergence for two discrete distribution $p$ and $q$ is defined as:

$$JS(p, q) = \frac{1}{2}KL(p||m) + \frac{1}{2}KL(q||m)$$

where $m = \frac{1}{2}(p + q)$ and $KL(p||m) = \sum_i p_i \log \frac{p_i}{m_i}$ is the Kullback-Leibler divergence. The similarity between two subjects $i$ and $j$ is then defined as:

$$\text{similarity}(i, j) = e^{-JS(\pi_i, \pi_j)}$$

where, with a little abuse of notation, $\pi_i$ and $\pi_j$ are the average mixture proportion for subject $i$ and $j$ respectively.

Figure 7 represents the similarity matrix obtained from the HDP (which automatically infers the number of activities) and the LDA (which pre-specifies the number of activities to 5). Visual inspection can easily see a clear separation in the result from the HDP compared to LDA.

To further substantiate and understand the proposed framework, we run LDA for various pre-defined values of $K$ and construct a gross ground truth by dividing 11 subjects into four groups according to the room they share: the first includes five PhD students, and the next six users are divided into 3 groups (two subjects in one room except subjects 7 and 11 who are in two adjacent rooms). We then use the Affinity Propagation (AP) algorithm [25] to perform a clustering. AP uses the similarity matrix as input and Figure 8 reports the clustering results using popular performance metrics in clustering and data mining literature which includes: F-measure, cluster parity, Rand index and normalized mutual information (NMI) (details can be found in [32, 41]). As it can be clearly seen, HDP achieves the best result without the need of specifying
Figure 7: Similarity matrix obtained for 11 subjects using colocation Bluetooth data from sociometric badge dataset. Similarity is computed for each pair of users using the Jensen-Shannon divergence between mixture proportions of activities being used by the two subjects.

Figure 8: Performance of LDA + AP and HDP + AP on colocation data. Since HDP automatically learns the number of activities (which is 5 in this case), we repeat this result for HDP over the horizontal axis for easy comparison with LDA.
Table 1: Visualization for nine latent activities learned by HDP, each is a discrete distribution over the set of $V = 69$ subjects in the Reality Mining data. Each number corresponds to an index of a subject and its size is proportional to the probability within that activity. Tag cloud column visualizes the tag cloud of affiliation labels for each activity.

the number of activity (topics) in advance whereas LDA results are sensitive and dictated by this model specification. HDP has achieved a reasonably high performance: an F-measure and NMI close to 0.9 and Rand index over 0.9.

4.2 Reality Mining data

Reality Mining [20] is a well-known mobile phone dataset collected at MIT Media Lab, involving 100 users over 9 months; a wide range of information from each phone was recorded (see the appendix of [37] for a full description of the database schema). To demonstrate the feasibility of our proposed framework, we focus on the Bluetooth events. Using Bluetooth recorded events as colocation information, our goal is to automatically learn activity of subject colocation in this dataset. Using this information, we further compute the similarity between subjects and group them into clusters. The affiliation information recorded in the database shall be used as the ground truth in which subjects can be roughly divided into labeled groups (e.g., staff, master frosh, professor, Sloan business school students and so on). Removing groups of less than 2 members and users whose affiliations are missing, we ended up with a total of $V = 69$ subjects
Figure 9: a) the matrix represents the probability of latent activities (topics) given subjects (69 users from Reality Mining data) and b) after applying a hard-threshold to convert it to a binary adjacency matrix; c) The graph visualization based on adjacency matrix inferred from latent activities with the node colors represent the affiliation labels from the ground truth.
divided into 8 groups in our dataset. For each subject, we construct one data point for every 10 minutes of recording. Each data point is a 69-dimensional vector whose element is set to 1 if the corresponding subject is present and 0 otherwise (self-presence is always set to one). All data points for each subject is treated as one collection of data, analogous to a document or a group in our HDP description.

Similar to colocation data from sociometric badges, each latent topic is a multinomial distribution over the set of users \(\{1, 2, \ldots, V = 69\}\). In this case, nine activities have been discovered and visualized in Table 1. Visual inspection reveals some interesting activities: activity 3, for example, shows a strong colocation of only two subjects 21 and 40; activity 9 shows that subject 55 mostly works alone; activity 2 shows a typical colocation of Sloan business school students and so on. Figure 9 attempts to construct a clustering structure directly using the mixture probability of latent activities assigned for each user. This colocation network is generated by thresholding the (topic, user) mixture matrix learned from HDP. The resultant matrix is then used as the adjacency matrix. In this network, the Sloan business school students and the master frosh students are usually colocated with other members of the same group. On the other hand, members of Media Lab might be colocated with members of other affiliation groups.

To further demonstrate the strength of the Bayesian nonparametric approach, we again perform a clustering experiment similar to the setting reported for colocation data in Section 4.1.2 and compared against LDA. We use the mixture proportion of activity and Jensen-Shannon divergence to compute the distance between users and similarity measure in Eq 33. Figure 10 shows the similarity matrix for Reality Mining data. It can be clearly seen that Sloan business school students and master frosh students are grouped into two separate groups. This is because the Sloan business school is physically adjacent to the Media Lab, while master frosh students only join the lab in some particular time. Table 2 shows ground truth groups in Reality Mining data and Table 3 presents the clustering result using HDP and AP. Figure 11 further

<table>
<thead>
<tr>
<th>Group</th>
<th>Group name</th>
<th>Subject ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1styear-grad</td>
<td>8 20 36 51</td>
</tr>
<tr>
<td>2</td>
<td>3rdyear-grad</td>
<td>32 66</td>
</tr>
<tr>
<td>3</td>
<td>masfrosh</td>
<td>2 19 21 40 42 47</td>
</tr>
<tr>
<td>4</td>
<td>ML-grad</td>
<td>1 7 16 31 37 43 44 46 60 64 67</td>
</tr>
<tr>
<td>5</td>
<td>new-grad</td>
<td>4 6 11 12 46 48 56 63 69</td>
</tr>
<tr>
<td>6</td>
<td>sloan</td>
<td>3 9 14 15 17 18 23 25 26 27 29 30 33 34 35 38 45 49 50 52 54 57 59 65</td>
</tr>
<tr>
<td>7</td>
<td>staff</td>
<td>10 24 62</td>
</tr>
<tr>
<td>8</td>
<td>student</td>
<td>5 13 22 28 39 41 53 55 58 61 68</td>
</tr>
</tbody>
</table>

Table 2: Eight ground truth groups in Reality Mining data.
Figure 10: Similarity matrix for 69 subjects from Reality Mining data using the results from HDP.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Subject ID</th>
<th>Tag cloud (group name)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 15</td>
<td>ML−grad student</td>
</tr>
<tr>
<td>2</td>
<td>5 8 10 11 13 22 24 28 39 41 51 56 58 61 68</td>
<td>student</td>
</tr>
<tr>
<td>3</td>
<td>3 9 14 15 17 18 23 25 26 27 29 30 33 34 35 38 45 49 50 52 54 59 65</td>
<td>sloan</td>
</tr>
<tr>
<td>4</td>
<td>7 12 32 36 46 63 66</td>
<td>3rd year−grad  new−grad</td>
</tr>
<tr>
<td>5</td>
<td>2 19 21 40 42 47</td>
<td>masfrosh</td>
</tr>
<tr>
<td>6</td>
<td>4 16 20 31 43 64 67 69</td>
<td>ML−grad new−grad</td>
</tr>
<tr>
<td>7</td>
<td>37 44 48</td>
<td>ML−grad new−grad</td>
</tr>
<tr>
<td>8</td>
<td>53 57</td>
<td>sloan student</td>
</tr>
<tr>
<td>9</td>
<td>6 60 62</td>
<td>ML−grad new−grad staff</td>
</tr>
</tbody>
</table>

Table 3: Clustering result data using HDP + AP.
Figure 11: Performance of LDA + AP and HDP + AP on Reality Mining data.

shows the performance comparison of HDP against LDA (run for various values of $K$). Again HDP consistently delivers a performance comparable to the best performance of LDA which requires searching over the parameter space.

5. Conclusion

Learning latent activities which are non-trivially present in the raw sensor data is important to many activity-aware applications and this chapter has explored the use of Bayesian nonparametric models for this task. In particular, we make use of the hierarchical Dirichlet process to discover movement and colocation activities from social signals. The key advantage of this approach is its ability to learn the number of latent activities automatically and grow with the data – thus completely bypassing the fundamental difficulty of model selection encountered in parametric models. We use data collected by sociometric badges in the lab as well as the public Reality Mining dataset. In all cases, we have demonstrated how the framework can be used, illustrated insights in the activities discovered and used them for clustering tasks. To demonstrate the strength and key advantages of the framework, we have rigorously compared the performance against state-of-the-art existing models and demonstrated that high performance is consistently achieved.
References


