Methodologies for Continuous Cellular Tower Data Analysis

Nathan Eagle[1,2], John A. Quinn[3], and Aaron Clauset[2]

[1] Massachusetts Institute of Technology, 20 Mass Ave, Cambridge, 02139
[2] The Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501
[3] Makerere University, Kampala, Uganda
nathan@mit.edu, john.quinn@ed.ac.uk, aaronc@santafe.edu

Abstract. This paper presents novel methodologies for the analysis of continuous cellular tower data from 215 randomly sampled subjects in a major urban city. We demonstrate the potential of existing community detection methodologies to identify salient locations based on the network generated by tower transitions. The tower groupings from these unsupervised clustering techniques are subsequently validated using data from Bluetooth beacons placed in the homes of the subjects. We then use these inferred locations as states within several dynamic Bayesian networks (DBNs) to predict dwell times within locations and each subject's subsequent movements with over 90% accuracy. We also introduce the X-Factor model, a DBN with a latent variable corresponding to abnormal behavior. By calculating the entropy of the learned X-Factor model parameters, we find there are individuals across demographics who have a wide range of routine in their daily behavior. We conclude with a description of extensions for this model, such as incorporating contextual and temporal variables already being logged by the phones.

1 Introduction

Every one of the approximately 4 billion mobile phones in use today have continuous access to information about proximate cellular towers. We believe these continuous cellular tower data streams can provide valuable insight into a user's behavior. Here we introduce a novel method of segmenting, validating and modeling this data. A major contribution of this paper involves the application and design of community structure algorithms that are appropriate for the identification of location clusters relevant to a user's life. We show that using temporal data from cellular towers, information every phone has access to, a simple generative model can be used to infer these salient locations and anticipate subsequent movements.

There has recently been a significant amount of research quantifying and modeling human behavior using data from mobile phones. We will highlight a selection of the literature on GSM trace analysis and subsequently discuss recent work on location segmentation and movement prediction from GPS data.

Mobile phones are continuously, passively monitoring signals from proximate cellular towers. However, due to power constraints, a mobile phone typically

H. Tokuda et al. (Eds.): Pervasive 2009, LNCS 5538, pp. 342–353, 2009.

does not continuously send back similar signals alerting the nearby towers of its particular location. While there has been recent work on analysis of call data records (CDR) from mobile phone operators [1,2], this data only provide estimates of locations when the phone is in use. Additionally, the only method of obtaining continuous cellular tower data without working with an operator is by installing a logging application on the mobile phone itself.

There have been a variety of projects that have involved installing a mobile phone application that logs visible cellular towers and Bluetooth devices on a set of subjects phones including HIIT's Context project, MIT's Reality Mining project [3] and the PlaceLab [4,5] research at Intel Research. Additionally, other research projects have demonstrated the utility of cellular tower data for a broad spectrum of applications ranging from contextual image tagging [6] to inferring the mobility and location of an individual [7,8,9]. Generally this logging software records between one to four of the cellular towers with the highest signal strength, however, recent research suggests it is possible to localize a handset down to 2.5 meter accuracies if the number of detected towers is dramatically increased [10].

Dynamic Bayesian Networks (DBNs) have been widely used for quantifying and predicting human behavior. For analysis of human movement, typically these models involve location coordinates that are much more precise than cellular tower data, such as GPS data. These models are trained on general human movement [11] or more specific data such as transportation routes [12].

As opposed to the previous work above, our dataset comes from randomly sampled individuals in a large US metropolitan city. We introduce several segmentation algorithms taken from the community structure literature and apply them to networks of cellular towers. Coupling bluetooth beacon data placed in the homes of each subject with the tower data, we validate the output of the community structure algorithms with the community of towers co-present with the beacon exposures. We then describe several DBNs that use the inferred locations clusters as states to parametrize and predict subsequent movements. One such DBN we use for behavioral modeling includes a latent variable, the X-Factor, corresponding to a binary switch indicative of "normal" or "abnormal" behavior. We compare the entropy of the learned X-Factor parameters across different demographics and conclude with ideas for extensions to these models as future work.

2 Methods

2.1 Data Description

Our data was generated from the phones of 215 subjects from a major US city. After providing informed consent, these subjects were given phones that logged the ID of the four cellular towers with the strongest signal strength every 30 seconds. Additionally, the phones conducted Bluetooth scans every minute. Bluetooth beacons were deployed in the homes of each subject; as the beacons are detected only if the phone is within 10 meters of the beacon, detection implies the subject is at home. The data was compressed on the mobile phone and uploaded to a central server after each day.

In contrast to previous datasets, every subject in our study was randomly sampled from a particular city. By offering a smartphone and free service, over 80 percent of the randomly selected individuals agreed to participate in the study. The demographic information we have about the subjects is evenly distributed among ethnic groups and income levels, accurately reflecting the distribution that makes up the city's inhabitants. No longer constrained to the study of academics or researchers, our data represents one of the first comprehensive behavioral depictions of the inhabitants within a major urban city.

2.2 Segmentation via Community Structure

Each phone records the four towers with the strongest signal at 30 second intervals. This data can therefore be represented as a cellular tower network (CTN) where each node is a unique cellular tower, an edge exists between two nodes if both towers co-occur in the same record, and each edge is weighted with the total amount of time (over all records) the pair co-occurred. A CTN is generated for each of the subjects, which includes every tower logged by the phone during the 5-month period. The nodes in the CTN that have the highest total edge weight (the node's "strength") correspond to the towers that are most often visible to the phone. Further, a group of nodes with a large amount of weight within the group, and less weight to other nodes, should correspond to a "location" where the user spends a significant amount of time. Figure 2 shows a 32-tower subgraph of one CTN, segmented into five such locations.

To allow for a meaningful comparison, we use three qualitatively different heuristics for clustering nodes into locations.

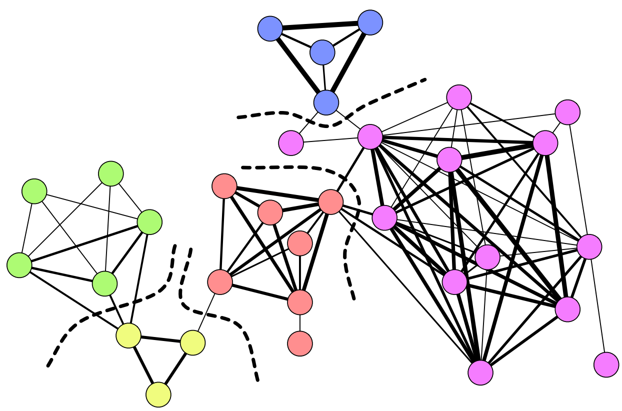

Fig. 1. A 32-tower subgraph of one of our cellular tower networks, segmented into five "locations," clusters of nodes in which towers frequently co-occur in the phone's records

Ncut. The first segmentation algorithm depends on Shi and Malik's *normalized cut* (Ncut) criterion [13], which, like many cut criteria, is NP-hard to optimize. Our implementation uses a spectral approach to find a bisection of the graph that minimizes the size of the normalized cut. Applied recursively, a graph can be split into a specified number of dense clusters. Although originally developed to segment images, the Ncut method can naturally be applied to networks.

Q-Modularity. The second method, drawn from the large literature on detecting "communities" in complex networks [14], depends on Newman and Girvan's popular *modularity* measure Q [15], which measures the density of clusters relative to a simple, randomized null model.

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right] \tag{1}$$

where l_s is the number of edges between the nodes within cluster s, L is the total number of edges in the network, and d_s is the sum of degrees of the nodes in cluster s. While finding the segmentation that maximizes Q is NP-complete, there has been a significant amount of work towards this goal. Although also NP-hard to maximize, we use Clauset *et al.*'s greedy optimizer [16], which has been shown to perform reasonably well on real-world data.

Threshold Groups. The third method is a simple-minded heuristic: we first identify the nodes in the upper decile of "strength," and then perform a breadth-first search on the induced subgraph. Each connected component in this subgraph is labeled as a unique location, and all remaining nodes in the original graph placed in an additional group.

Although all based on somewhat similar principles, in practice these methods produce dramatically different segmentations of our CTNs. This is in part because the first algorithm requires as input the number of segments to be found, unlike the other two.

2.3 Inference via Bluetooth Beacons

One objective measure of these clusterings is to use independent information derived from the Bluetooth beacons, installed in the homes of each subject in the study. Every minute the phone scans for visible Bluetooth devices and if a beacon is within 10 meters of the phone, it is logged as proximate. Creating training data from the set of cellular towers detected at the same time as the bluetooth beacons, we have used several methodologies to infer if a subject is at home given a particular set of visible cellular towers.

Bayesian Posteriors. It is possible to calculate the posterior probability a subject is home, $P(L_{home})$, conditioned on the four towers currently detected by the phone, T_{abcd}, using the likelihood, the marginal and the prior probability of being at home (based on the beacon data).

$$P(L_{home}|T_{abcd}) = \frac{P(T_{abcd}|L_{home})P(L_{home})}{P(T_{abcd})} \tag{2}$$

Gaussian Processes. While the naive Bayesian model above works well in many cases, simply using the ratio of tower counts co-present with the Bluetooth beacon tends to fail if the phone regularly moves beyond ten meters of the beacon while still staying inside the home. Instead of normalizing by total number of times each tower is detected, it is possible to obtain additional accuracy by incorporating the signal strengths from the detected towers. There are many models for signal strength of a single cellular tower, t. $p_t(s_t|\mathbf{l})$, one such model uses training data to estimate Gaussian distributions over functions modeling signal propagation from cellular towers [17]. In our case, the training data comes from the signals of towers detected at the same time as the Bluetooth beacon in the subject's home, and the inference is binary (home or not home); however, these models are easily extendable for more broad localization.

Deviations in Tower Signal Distributions. The two models above generate a probability of being at home associated with a single sample of detected towers (ie: the four tower IDs and their respective signal strengths). However, during the times when a subject is stationary, the phone continuously collects samples of the detected towers' signal strengths. These samples can form 'fingerprint' distributions of the expected signal strengths associated with that particular location. It is possible to detect deviations within these distributions of signal strengths using a pairwise analysis of variance (ANOVA) with the Bonferroni adjustment to correct for different sample sizes. Training the home distributions on the times when the beacon is visible (or if there are no beacons, on times when the subject is likely home such as 2-4am), an ANOVA comparing this home distribution with a distribution of recent tower signal strengths makes it possible to identify if the subject is truly at home, or is at a next-door neighbor's house. In previous work, such tower probability density functions have successfully localized a phone down to the office-level [3].

2.4 Prediction via Dynamic Bayesian Networks

The clusters of towers identified above can be incorporated as states of a dynamical model. Given a sequence of locations visited by a subject, we can learn patterns in their behaviour and calculate the probability of them moving to different future locations. We start with a baseline dynamical model and introduce additional observed and latent variables in order to model the situation more accurately.

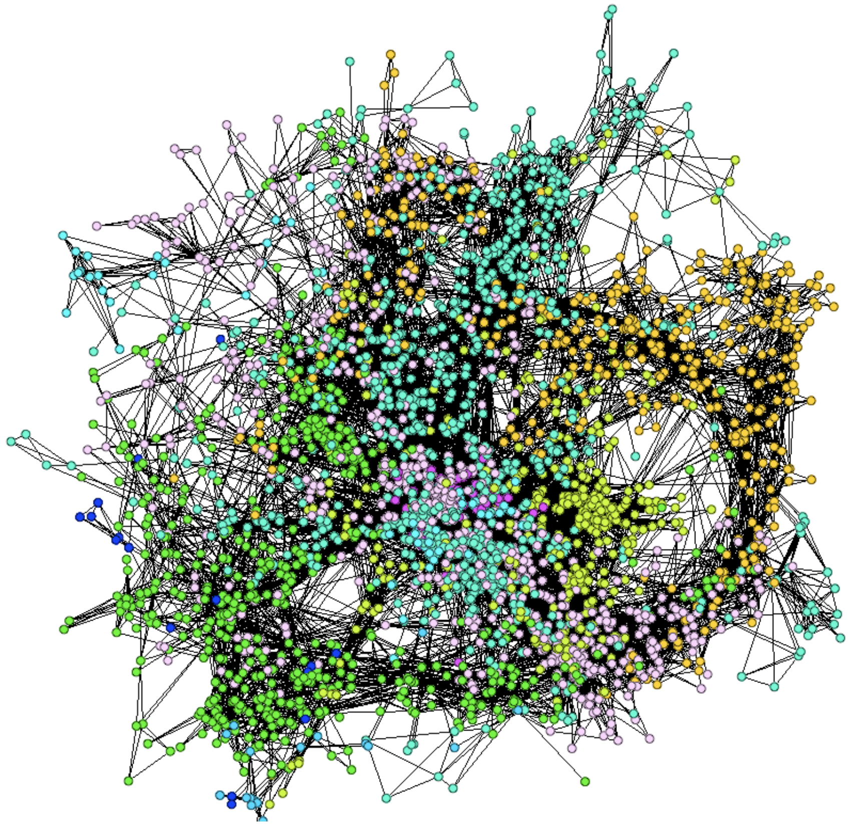

Fig. 2. CTN Segmentation. The giant component of a subject cellular tower network, segmented into 15 major location clusters (represented by 15 colors) using the Q-modularity community structure method.

The simplest dynamical Bayesian network we can use for location prediction is a Markov chain, in which the location y_t depends only on the location at the previous time step, y_{t-1}. The maximum likelihood transition probabilities $p(y_t|y_{t-1})$ can easily be estimated. Given evidence that a user is in a particular location at time t, this allows us to calculate the τ-step-ahead prediction $p(y_{t+\tau}|y_t)$.

We note that patterns of movement in practice are dependent on the time of day and the day of week. Subjects typically exhibit different dynamics on weekday mornings than on Saturday evenings, for example. Figure 3(a) shows an extended model where the probability of being in a location is also dependent on the hour of day h_t and the day of week d_t. In the experiments below, we code h_t to take on the values "morning", "afternoon", "evening" and "night", and code d_t to take on the values "weekday" or "weekend". After learning maximum likelihood parameters we can calculate the predicted density $p(y_{t+\tau}|y_t, d_{t+1:t+\tau}, h_{t+1:t+\tau})$ for new observations from the same user.

2.5 X-Factors for Abnormality Modeling

While there is strong structure in human behavior, there are also regular deviations from the standard routines. We incorporate an additional latent variable into our model to quantify the variation in behavior previously unaccounted for in the fully observed models above.

The model we use for this is shown in Figure 3(b). Here we factorize the location variable so that it depends on a hidden "abnormality" variable a_t. The model can now switch between "normal" and "abnormal" behaviour depending on whether a_t is 0 or 1 respectively, as demonstrated in previous physiological condition monitoring work [18].

We expect abnormal dynamics to be related to the normal dynamics but with a broader distribution. When estimating these dynamics, we therefore want to keep relevant structure in the dynamics (e.g. transitions between physically neighboring locations are still more likely), while allowing wider possibilities including non-zero probability of transitions not seen in the training data. We can achieve this effect by tying the parameters between the normal and abnormal transition probabilities such that $p(y_t|y_{t-1}, d_t, h_t, a_t = 1)$ are a smoothed version of $p(y_t|y_{t-1}, d_t, h_t, a_t = 0)$. To smooth the transition matrices for every combination of d_t and h_t we add a small constant ξ to each entry in the matrix and renormalize.

Learning of this model can be done with expectation-maximization (EM). We perform a standard E-step to calculate the probability of being in the normal or abnormal regime at each time frame, then modify the standard M-step to use the parameter tying above. In the experiments below, we set $\xi = .1$ by hand, though in principle this parameter can also be learnt using EM. Increasing ξ effectively specifies that a sequence has to depart further from normal dynamics in order to be considered "abnormal".

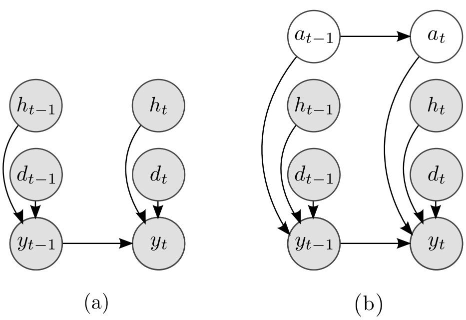

Fig. 3. Two DBN models used for location prediction. Shaded nodes are observed and unshaded nodes are latent; y_t denotes location, d_t denotes day of week, h_t denotes hour of day, and a_t denotes abnormal behaviour (all at time t). Panel (a) shows a fully observed model as a contextual Markov chain (CMC), and panel (b) shows the X-factor model, where location is additionally conditioned on the latent abnormality variable.

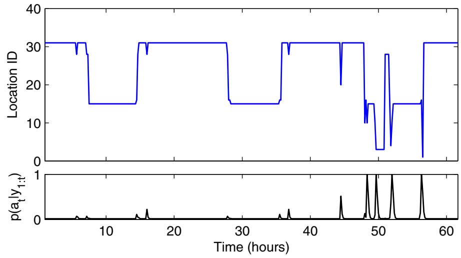

Fig. 4. Inferred points of abnormality using the X-Factor model. Each weekday the subject moves consistently between home (location 31) and work (location 15), but on the third day makes some extra, unusual journeys. The locations in this example were given by the Group Threshold segmentation method.

3 Results and Discussion

3.1 Segmentation Validation

We have shown how data collected from installed Bluetooth beacons can be used to create a known cluster of towers associated with each subject's home. We used this known cluster to validate each segmentation algorithm, selecting twenty locations for the Ncuts technique. Table 1 categorizes the community detecition algorithms by how well they detected the "home" towers as defined by the Bluetooth beacons, C_{BT}. The home cluster of towers generated by the Threshold Groups technique incorporated C_{BT} for every subject, $P(C_{BT} \subset C_H) = 100\%$, while this was the case for the Q-Modularity technique only 86% of the time. However, the other important statistic is the ratio of the number of the Bluetooth home towers, $N_{C_{BT}}$, to the number of towers in the inferred home cluster, N_{C_H}. This ratio describes how many additional towers were included in the inferred home location; for example, the Q-Modularity home cluster has a ratio of .18, indicating that its home cluster contains approximately five times as many towers as needed. Despite averaging the most number of clusters, the Ncuts home cluster has a ratio of .0061, implying that a few large clusters tend to dominate these segmentations.

3.2 Dwell and Movement Prediction

The three DBNs described above were trained on sequences of transitions between the locations that were inferred by each segmentation method. To compensate for the bias towards self-transitioning (at virtually every instance, the most likely event will be that the subject does not change locations), we compare the models success only on instances when a subject is about to transition between inferred locations. The DBNs are tasked with predicting the location where the subject is about to move. Table2 lists these prediction accuracies for the three segmentation methods and the two full-observed Markov models. While

Table 1. Segmentation Validation via Bluetooth Beacons. μ_{N_C} is the average number of clusters generated by each segmentation method. $P(C_{BT} \subset C_H)$ represents the probability that the set cellular towers associated the Bluetooth beacon at the subject's home, C_{BT}, is fully contained in a single cluster, C_H. The last column corresponds to the ratio of the actual number of home towers, $N_{C_{BT}}$ to the number of home towers inferred by the different segmentation methods, N_{C_H}. A small number corresponds to incorporating a large number of towers within the home cluster.

method	μ_{N_C} (σ)	$P(C_{BT} \subset C_H)$	$\frac{N_{C_{BT}}}{N_{C_H}}$
Ncuts	20 (0)	.93	.0061
Q-Modularity	13.3 (11.7)	.86	.18
Threshold Groups	6.8 (13.7)	1.0	.045

Table 2. Transition Accuracy and Dwell Errors. For every instance a subject moves between two clusters of towers, the DBN can be used to predict the subsequent cluster. The different accuracies between the segmentation methods are due to not only how well the clustering techniques performed at identifying the true salient locations, but also to the number and size of the clusters (described in Table 1). Given these high accuracies, the inclusion of the temporal observations in the Contextual Markov Chain (CMC) does not appear to provide significant improvement to the standard Markov chain (MC).

method	MC Transition Prediction	CMC Transition Prediction	MC Dwell Error (minutes)	CMC Dwell Error (minutes)
Ncuts	.932	.933	79.1	78.9
Q-Modularity	.953	.954	91.0	75.7
Threshold Groups	.992	.992	89.2	84.1

the X-factor model provides additional information about the regularity of a particular behavior, its accuracy is identical to the contextualized Markov model and was not included in the table. It is of interest that the highest accuracies did not come from the segmentation methods that provided the largest cluster sizes (Ncuts), but rather the smallest number of clusters (Threshold Groups). However, a direct comparison between these accuracies is not possible due to the differences in the dimensionality of the state spaces. A model with fewer inferred locations (N_C) should be expected to do better because it has less potential for a wrong prediction. In the extreme, a model with a single state will always be correct, yet obviously adds little value. Therefore, while the Threshold Groups segmentation method, with an average of 6.8 inferred salient locations ($\sigma = 13.7$), generated accuracies of over 99%, future work in predicting location dwell times may provide more conclusive information about the dominance of one particular segmentation method over the others. Given the extremely high accuracies using an unconditioned Markov model, incorporating information about the time of day and day of the week unsurprisingly adds little additional value.

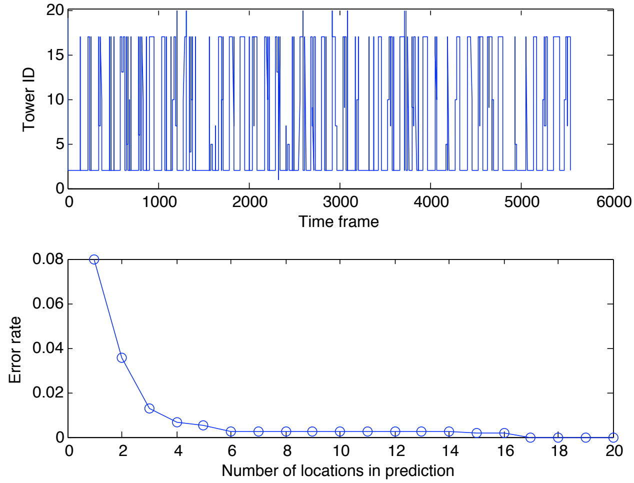

Fig. 5. A sequence of transitions between towers (top) and the average error rates for predicted transitions (bottom). The X-factor model was tested on approximately one month of movement segmented using Ncuts into 20 locations. While the top inferred location is 92% correct for this set of data, the subsequent location is in the top four locations over 99% of the time.

3.3 Entropic Individuals

By calculating the standard Shannon information entropy metric of the learned transition probabilities of the X-factor model, $H = -\sum p \times log_2(p)$, we are able to quantify the amount of behavioral regularity of each subject. The means and variances of this entropy metric are segmented across demographics in Table 3. Of particular note is the high entropy variance, indicating that there are individuals across all demographics whose behavioral patterns are seemingly unstructured. This finding runs contrary to previous research conducted on university students and staff which suggested behavioral entropy is correlated with demographics [3].

3.4 Future Work

This paper has provided the groundwork for the design of increasingly sophisticated models based on data from mobile phones that incorporate contextual and temporal variables and can use demographic priors for bootstrapping. For example, if the discovered Bluetooth devices can be clustered based on co-presense, it may be possible to classify particular Bluetooth phones as family, colleagues, and friends, incorporating the proximity of these individuals as observational variables. Additionally, the phones in this study also sample the ambient audio environment periodically to detect the subjects' media consumption, information that should also make for an intriguing additional observed variable in the DBN. Lastly,

Table 3. Demographic Entropy. The entropy of the conditional probability table from the X-factor model using the Group Threshold method was averaged across demographics. The results show extremely high variance, with entropic individuals in virtually every demographic as well as subjects with significant structure in their daily behavior.

demographic (N)	$\mu_{entropy}$ ($\sigma \times 10^2$)
Age:	
under 35 (107)	30.1 (4.2)
35 and over (108)	28.0 (4.2)
Gender:	
Male (136)	28.3 (4.4)
Female (79)	30.3 (3.8)
Income:	
over $60,000 (73)	34.2 (4.3)
$60,000 and under (140)	26.4 (4.0)
Education:	
College Grad (79)	31.2 (4.3)
No College Degree (125)	27.7 (4.1)

we would like to explore the potential of using demographic bootstrapping to aid in efficient model parameterization as introduced in similar models [12].

We have demonstrated the potential to repurpose algorithms developed originally to quantify community structure within graphs to identify salient locations within a cellular tower network. We have validated these unsupervised clustering algorithms on a known cluster of towers using the Bluetooth beacon installed in each of our randomly sampled subjects' homes. The resultant set of inferred clusters of towers correspond to salient locations and are incorporated as states in our DBN models. We introduced the X-Factor model to detect behaviors that deviate from a given routine by incorporating an additional latent variable corresponding a normal / abnormal switch. By calculating the entropy of the transition matrix from this model we were able to quantify the amount of structure in the daily routines of different demographics. It is our hope that these analytical methodologies will provide a framework for future studies of this rich behavioral data, currently being generated by the majority of humans today.

References

1. González, M.C., Hidalgo, C.A., Barabási, A.L.: Understanding individual human mobility patterns. Nature 453(7196), 779–782 (2008)
2. Onnela, J., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., Kertesz, J., Barabasi, A.L.: Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences 104, 7332–7336 (2007)
3. Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10, 255–268 (2006)

4. Chen, M., Sohn, T., Chmelev, D., Haehnel, D., Hightower, J., Hughes, J., LaMarca, A., Potter, F., Smith, I., Varshavsky, A.: Practical metropolitan-scale positioning for gsm phones. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 225–242. Springer, Heidelberg (2006)
5. LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J.: Place lab: Device positioning using radio beacons in the wild. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) Pervasive 2005. LNCS, vol. 3468, pp. 116–133. Springer, Heidelberg (2005)
6. Davis, M., King, S., Good, N., Sarvas, R.: From context to content: leveraging context to infer media metadata. In: Proceedings of the 12th annual ACM international conference on Multimedia, New York, NY, USA, October 10-16 (2004)
7. Sohn, T., Varshavsky, A., LaMarca, A., Chen, M.: Mobility detection using everyday gsm traces. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 212–224. Springer, Heidelberg (2006)
8. Laasonen, K., Raento, M., Toivonen, H.: Adaptive on-device location recognition. In: Ferscha, A., Mattern, F. (eds.) Pervasive 2004. LNCS, vol. 3001, pp. 287–304. Springer, Heidelberg (2004)
9. Hightower, J., Consolvo, S., LaMarca, A., Smith, I.: Learning and recognizing the places we go. In: Beigl, M., Intille, S.S., Rekimoto, J., Tokuda, H. (eds.) UbiComp 2005. LNCS, vol. 3660, pp. 159–176. Springer, Heidelberg (2005)
10. Otsason, V., Varshavsky, A., LaMarca, A., de Lara, E.: Accurate gsm indoor localization. In: Beigl, M., Intille, S.S., Rekimoto, J., Tokuda, H. (eds.) UbiComp 2005. LNCS, vol. 3660, pp. 141–158. Springer, Heidelberg (2005)
11. Ashbrook, D., Starner, T.: Using gps to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing 7, 275–286 (2003)
12. Liao, L., Patterson, D., Fox, D., Kautz, H.: Learning and inferring transportation routines. In: Proceedings of the Nineteenth National Conference on Artificial Intelligence, pp. 348–353 (January 2004)
13. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Learning 22(8), 888–905 (2000)
14. Newman, M.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences (January 2006)
15. Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69 (January 2004)
16. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Physical Review E 70(6) (December 2004)
17. Schwaighofer, A., Grigoras, M., Tresp, V., Hoffmann, C.: Gpps: A gaussian process positioning system for cellular networks. Advances in Neural Information Processing Systems 16 (January 2004)
18. Quinn, J., Williams, C., McIntosh, N.: Factorial switching linear dynamical systems applied to physiological condition monitoring. IEEE Transactions on Pattern Analysis and Machine Intelligence (January 2008)