Mobile Sensing for Data-Driven Mobility Modeling

Kashif Zia
Faculty of Computing and Information Technology
Sohar University, Oman

Katayoun Farrahi
Department of Computing
Goldsmiths, University of London, UK

Arshad Muhammad
Faculty of Computing and Information Technology
Sohar University, Oman

Dinesh Kumar Saini
Faculty of Computing and Information Technology
Sohar University, Oman

***Abstract*—The use of mobile sensed location data for realistic human track generation is privacy sensitive. People are unlikely to share their private mobile phone data if their tracks were to be simulated. However, the ability to realistically generate human mobility in computer simulations is critical for advances in many domains, including urban planning, emergency handling, and epidemiology studies. In this paper, we present a data-driven mobility model to generate human spatial and temporal movement patterns on a real map applied to an agent based setting. We address the privacy aspect by considering collective participant transitions between semantic locations, defined in a privacy preserving way. Our modeling approach considers three cases which decreasingly use real data to assess the value in generating realistic mobility, considering data of 89 participants over 6079 days. First, we consider a dynamic case which uses data on a half-hourly basis. Second, we consider a data-driven case without time of day dynamics. Finally, we consider a homogeneous case where the transitions between locations are uniform, random, and not data-driven. Overall, we find the dynamic data-driven case best generates the semantic transitions of previously unseen participant data.**

***Keywords*—mobile sensing; data-driven mobility model; agent based models**

I. Introduction

Large-scale mobile phone data for human behavior understanding has gained much popularity. Reality mining data has shown to be a useful tool in many scientific domains, including healthcare, and the social sciences. Here we consider the use of mobile data in the context of agent based models. One fundamental building block of any agent based system is mobility; what is the best way to generate agent mobility on a real physical space in a realistic manner. This achievement is critical for the successful use of agent based models in many interdisciplinary domains. Many current mobility models for agents are based on simple, homogeneous random processes. In this paper, we propose to use real human mobility data obtained by mobile phones to address this issue. A data-driven approach, particularly based on mobile sensing, has the advantage of offering realistic human tracks and time-varying dynamics, over many spectrums of the population, with differing possible timescales and sensors. A mobile data-driven approach is easily extendable to any geographical location as people ubiquitously carry their mobile phones everywhere. A collective approach, whereby the collective dataset is used for modeling, has the advantage of protecting individual participant privacy.

There has been few previous effort to develop mobility models from real traces; most previous works have focused on different data sources, such as wireless network data, survey data, and social interaction data for mobility modeling. An overview of simulation of traffic and pedestrians is presented in [1]. More realted to this work, a survey of data-driven perdestrians mobility models is given in [2]. A large portion of research in mobility is done in wireless networks [3], [4]. In [5], a hybrid mobility model is developed based on count data collected over a given campus map (number of people passing through various hallways on campus) and is not easily generalizable. In [6], real user traces are obtained by participant survey data, where 268 students are asked to keep a diary of their movements on campus. The participants record their locations, given five pre-defined types (classroom, library, cafeteria, off-campus, other), over one month. A Weighted Way Point (WWP) mobility model is proposed, and the focus of the model is towards destination selection. In [7], an ad hoc mobility model is defined and is founded on social relationship modeling, focusing on pairwise interactions as opposed to locations. In [8], mobile phone interaction patterns have been characterized based on relationships between participants. While all of these previous works consider the modeling of mobility, none of them are focused on mobile phone cell tower connection data, nor has the focus been on mobility modeling for synthetic mobility generation, for example in an agent based setting.

Previous work in the agent based community has addressed the issue of mobility from two differing points of view, density and crowding [9], [10], and movement between origin to destination [11], [12], [13]. Our work falls in the latter category, which is applicable to larger areas and more general scenarios. However, to the best of our knowledge, this is the first mobile phone location-driven approach for agent mobility simulation. In [11] an outdoor pedestrian mobility model is defined, where mathematical emulations to more realistic movements have been made. The model is not data-driven, and is not applicable to large-scale mobile sensed data.

The closest related work to ours is by Kim et al. [14] who develop a mobility model based on wireless network data. The goal of the work is to determine the real user *tracks* of the participants given their WiFi network traces over time. This work focuses on determining the accurate tracks taken by the participants given their access point coordinate sequence information. While this work falls under the category of data-

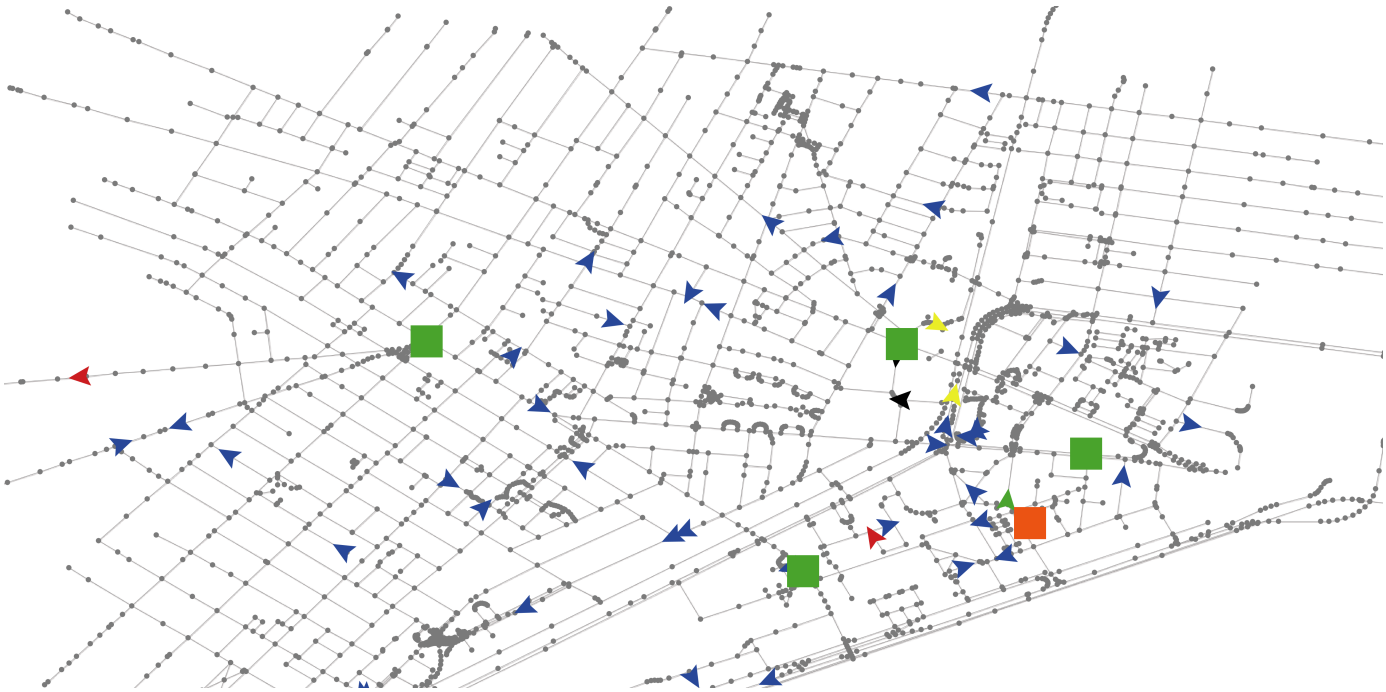

Fig. 1. Map of the physical space used for agent based simulation. The squares are the points of interest, the media lab in orange and the subway stops in green. The agents are marked as triangle and the arrows indicate the direction of motion. Different colors represent the different transitions, for example red is out to work, yellow is out to home, green is work to home, black is home to work, and blue is any location to out.

driven mobility modeling, the goal of the mobility model is very different. In [14], the goal of the model is to determine accurate user tracks based on network traces, though it is also used for generating synthetic mobility tracks. However, in this work, we consider agent mobility on a real map of the area in which the data is collected, and we consider location data obtained by mobile phones which provides a more general means for future mobility modeling and extensions of our approach.

In this paper, we present a basic framework to incorporate real mobile phone location data into an agent based simulation framework. To the best of our knowledge, it is the first mobile phone data-driven mobility model simulated in an agent based setting. Considering the cell tower connection data of 89 participants over a period of 6079 days, we consider their collective overall movements between three semantic locations, home, work, and out. Considering a real map of the Cambridge, Massachusetts area, where the mobile phone data was collected, we generate synthetic agent tracks which are then evaluated against mobile phone human location observations on unseen data. We compare three different settings, a data-driven approach with half-hourly dynamics, a data-driven approach without daily time dynamics but considering the average overall daily location information, and an approach without real data using a random assignment to location. Overall, we find using the time-varying transition probabilities of location results in the generation of agent mobility patterns which more closely approximates the location occurrences in the real data (considering a previously unseen test set).

II. Data-Driven Mobility Model Framework

A. Dataset

We consider the mobile phone cell tower connections obtained by the publicly available Reality Mining dataset [15]. The mobile phone data of 97 participants over 491 days is available. However, due to the large amount of noise (missing data for many days, missing semantic labels for location) we only consider days which contained a minimum of 20 hours of semantic location information, resulting in a total of 6079 days for evaluation. These 6079 days are from 89 of the participants in the dataset.

B. Space Model

The physical space in which agent movement takes place is taken from a neighbourhood in Cambridge, Massachusetts on the MIT campus. Openstreetmap [16] is used to obtain the shape files used in our space model to simulate the real physical space of MIT campus, shown in figure 1.

C. Points of Interest

We consider three semantic labels of places a person commonly visits: *home, work*, and *out*. A day is considered to be constructed of 48 location labels, which are the most often occurring location for the half-hour interval of the day. The physical locations of the cell towers are unknown, however, some participants labeled work and home locations. The participant labels are used to annotate the mobile phone data. In order to mark the points of interest geographically on a map, we consider several known landmarks. The MIT Media Lab, for which most of the participants are students and staff, is marked as the work location. We consider one work location at the moment, but this can be extended for

future simulations to consider more general regions of interest. All of the subway stations on the map are marked as home locations. The reasoning is, no matter where in the city the participants may live, they are very likely to travel to their homes by using the subways from campus. While there is some error in this reasoning, some participants may live on campus or choose other means to travel home, for the most part this assumption is true and it is a novel way to consider participants' homes without considering their privacy sensitive precise home information. Out locations are considered to be anywhere other than home and work.

D. Model Dynamics

We consider three scenarios for simulation, a data-driven approach with half-hourly dynamics (abbreviated as DD), a data-driven approach without dynamics (abbreviated as DN), and a random approach without data (abbreviated as R). The DD approach is based on the half-hourly region transition matrix (defined next in Mobility Model) averaged over the entire training data for the given time interval and therefore considers the daily time-evolving dynamics in agent location transitions. The DN approach uses the overall average daily region transition matrix obtained on the training data, and therefore does not consider the daily time dynamics but the overall daily average. The R approach considers a uniform random transition matrix, where the transition from regions are equi-probable.

E. Mobility Model

Our agent based model is simulated using NetLogo [17]. An $A = N \times N \times T$ region transition matrix is generated from the real data transition information between semantic locations. In the DD case, $N = 3$, corresponding to the semantic locations of home, work and out. $T = 48$, corresponding to the number of 30 minute intervals in a day. We do not differentiate between day types (for example, day of week), however, we model the dynamic behavior over the day. In the DN case, $T = 1$ and the region transition matrix is computed as an overall daily average. In the R case, $T = 1$ and the probability of region transition for every region is simply $1/N$.

We consider a set of 100 agents, initially distributed randomly between the home locations. Agents remember their home locations and always return to the same homes. Agents can be either stationary or mobile. At every 30 minute interval, the agents' next destination is sampled from A. If there is no change in state, the agent remains stationary, otherwise it departs towards the next destination sampled. The agent's next destination is chosen based on the probabilities in the region transition matrix. In the case of out, the agent can move towards any point on the map, chosen randomly. In the case of home, the agent moves towards her predefined subway station. In the case of work, the agent moves towards the MIT Media Lab. In figure 1, agents are illustrated with directed triangles, indicating the direction of movement. The different agent colors correspond to their region transitions, for example, red is out to work, yellow is out to home, green is work to home, black is home to work, and blue is any location to out.

F. Speed

Agents can have a maximum possible speed of 10 kph, however, this measure can be easily adjusted. There is a variation in speed across agents, which is determined randomly. The variation in speed can be up to 20% of the current speed of an agent.

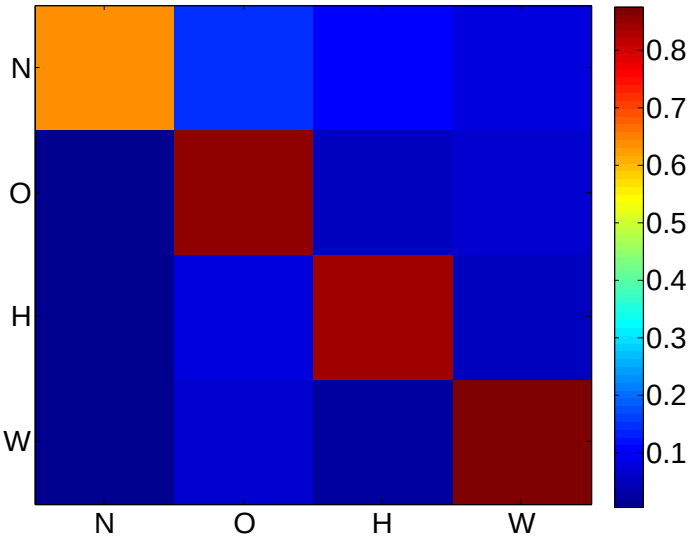

Fig. 2. The overall average region transition matrix A for the DN case, where the additional label N corresponds to "no data". The labels O, H, W correspond to out, home, work, respectively. The legend indicates the probability of the transition from on location to another, averaged over all days. Note, A does not include transitions to and from N, though it is shown for completeness.

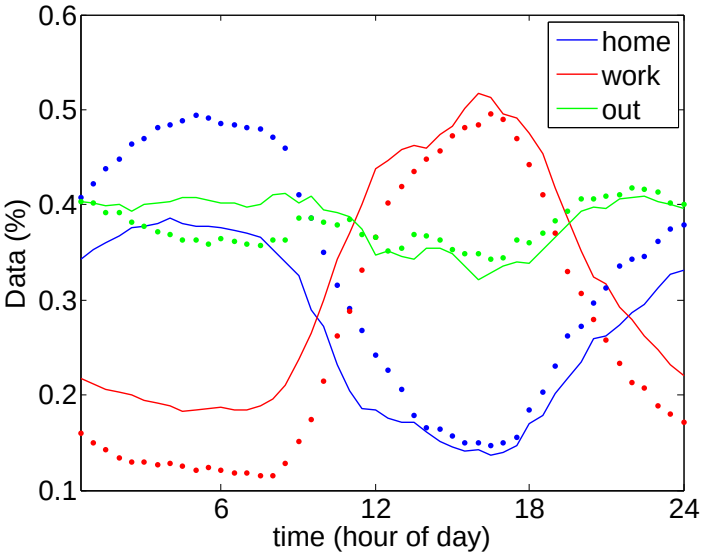

Fig. 3. Visualization of all the data over time. The solid lines represent the data used for training and the dotted lines represent the test data used for evaluation.

III. Simulation Results

A. Data

Every simulation result presented is run over 10 random simulations of the mobility model. For simulations, we divide the data into two partitions, a training set and a test set. The training set contains 50% of the days, randomly selected, and is used to generate the region transition matrix A. The remaining

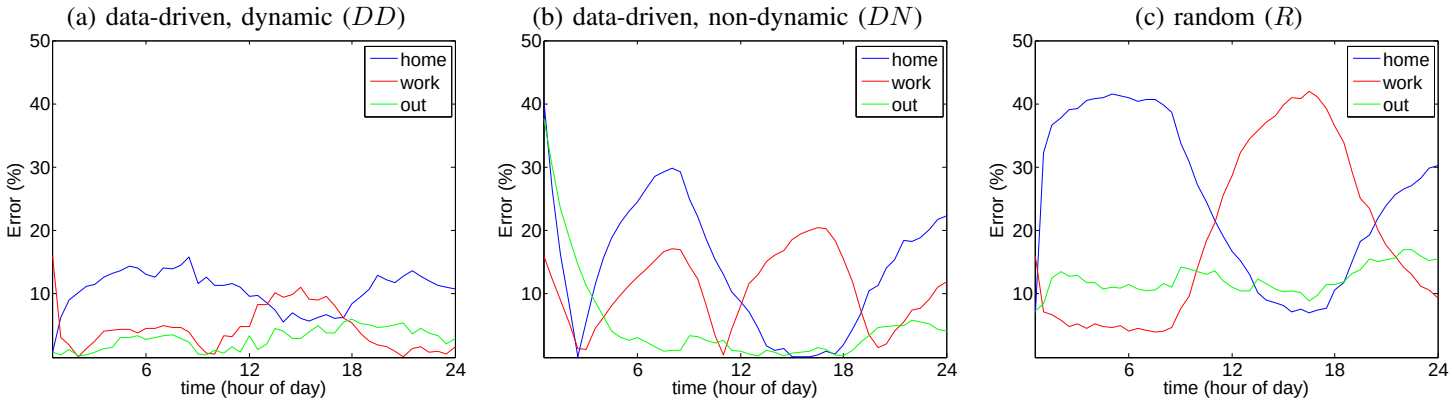

Fig. 4. The percentage of error computed over each half hour interval of the day is presented for the three cases, (a) the data-driven dynamic case, (b) the data-driven, non-dynamic case, and (c) the random case. The overall percentage of error is much higher in the random case (almost double).

50% of the days are used for evaluation. Note, the partitions are created in order to evaluate the generative ability of the framework on previously unseen data.

The data used for experiments is plotted in figure 3. For each half hour interval in a day, we plot the total average percentage of each location. Note, there are many sources of noise in the data, and there are often missing location labels, which is why the sum over the percentage of labels is never exactly 1. The solid lines show the training set and the dotted lines show the test set.

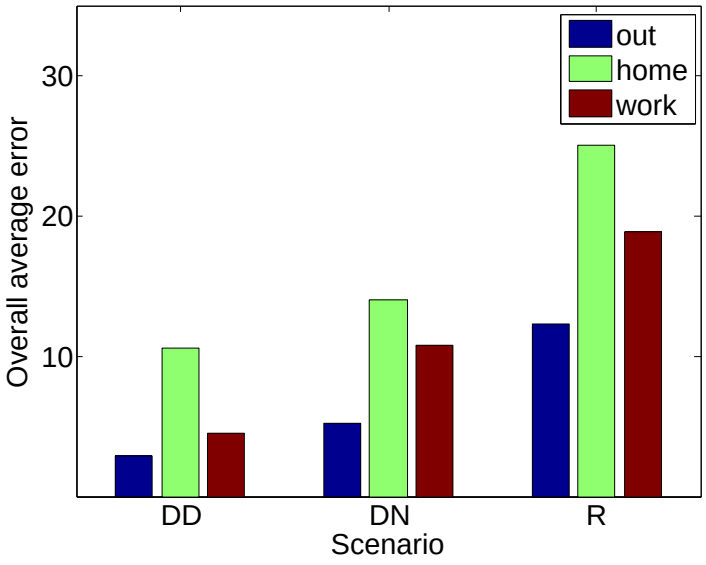

Fig. 5. The overall averaged error for the three cases. It is apparent that for every location type (home, work, out), the more real data is used (including dynamics versus no dynamics), the less the error.

B. Discussion

The agent based model is simulated in the 3 scenarios defined. Each simulation result is generated over the course of 1 day (averaged over 10 runs), where every 30 minutes the agents current locations are logged for evaluation. These results are compared to the test set. The error is the average total percentage of agents (participants) located at home, work, out computed as the absolute difference with the test set. The results in figure 4 are over time of day. In figure 4 (a), there is an overall least amount of error, particularly for the out location. The highest error occurs in all cases for the home location, particularly in the morning. In figure 5, the overall average error is plot for the three cases. The DN case performs better than the R case, and the DD case performs better than the DN case, showing the more data-driven information used, the better the agent mobility tracks mimic the real data.

IV. Conclusion

This work presents a data-driven mobility modeling framework where semantic locations obtained by mobile phone cell tower connection data are collectively used to formulate a mobility model. While the mobility model itself is simple, it is an initial component of our data-driven methodology for simulating agent mobility. Future work will explore more advanced techniques to incorporate the real location data into the framework. Machine learning tools, such as hidden markov models, will be the natural next step to consider for modeling. We will also further consider new sources for data-driven behavior modeling from mobile phone sensors, particularly GPS and Bluetooth physical proximity data.

References

[1] E. Papadimitriou, J.-M. Auberlet, G. Yannis, and S. Lassarre, "Simulation of pedestrians and motorised traffic: existing research and future challenges," *International Journal of Interdisciplinary Telecommunications and Networking (IJITN)*, vol. 6, no. 1, pp. 57–73, 2014.

[2] A. Hess, K. A. Hummel, W. N. Gansterer, and G. Haring, "Data-driven human mobility modeling: A survey and engineering guidance for mobile networking," *ACM Computing Surveys (CSUR)*, vol. 48, no. 3, p. 38, 2016.

[3] M. Al-Ayyoub, G. Husari, and W. Mardini, "Improving vertical handoffs using mobility prediction," *International Journal of Advanced Computer Science & Applications*, vol. 1, no. 7, pp. 413–419.

[4] M. B. Yassein, "Flying ad-hoc networks: Routing protocols, mobility models, issues," *International Journal of Advanced Computer Science & Applications*, vol. 1, no. 7, pp. 162–168, 2016.

[5] D. Bhattacharjee, A. Rao, C. Shah, M. Shah, and A. Helmy, "Empirical modeling of campus-wide pedestrian mobility observations on the usc campus," in *Vehicular Technology Conference, 2004. VTC2004-Fall. 2004 IEEE 60th*, vol. 4. IEEE, 2004, pp. 2887–2891.

[6] W.-j. Hsu, K. Merchant, H.-w. Shu, C.-h. Hsu, and A. Helmy, "Weighted waypoint mobility model and its impact on ad hoc networks," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 9, no. 1, pp. 59–63, 2005.

[7] M. Musolesi, S. Hailes, and C. Mascolo, "An ad hoc mobility model founded on social network theory," in *Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*. ACM, 2004, pp. 20–24.

[8] K. Farrahi, R. Emonet, and A. Ferscha, "Socio-technical network analysis from wearable interactions," in *International Symposium on Wearable Computers (ISWC)*, June 2012.

[9] D. Helbing, "Traffic and related self-driven many-particle systems," *Reviews of modern physics*, vol. 73, no. 4, p. 1067, 2001.

[10] K. Zia, A. Ferscha, A. Riener, M. Wirz, D. Roggen, K. Kloch, and P. Lukowicz, "Scenario based modeling for very large scale simulations," in *Distributed Simulation and Real Time Applications (DS-RT), 2010 IEEE/ACM 14th International Symposium on*. IEEE, 2010, pp. 103–110.

[11] R. Vogt, I. Nikolaidis, and P. Gburzynski, "A realistic outdoor urban pedestrian mobility model," *Simulation Modelling Practice and Theory*, vol. 26, pp. 113–134, 2012.

[12] J. Dijkstra, J. Jessurun, and H. J. Timmermans, "A multi-agent cellular automata model of pedestrian movement," *Pedestrian and evacuation dynamics*, pp. 173–181, 2001.

[13] K. Zia and A. Ferscha, "A simulation study of exit choice based on effective throughput of an exit area in a multi-exit evacuation situation," in *Proceedings of the 2009 13th IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications*. IEEE Computer Society, 2009, pp. 235–238.

[14] M. Kim and D. Kotz, "Extracting a mobility model from real user traces," in *In Proceedings of IEEE INFOCOM*, 2006.

[15] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, vol. 10, no. 4, pp. 255–268, 2006.

[16] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *Pervasive Computing, IEEE*, vol. 7, no. 4, pp. 12–18, 2008.

[17] U. Wilensky, "{NetLogo}," 1999.