Trajectory Simulation in Communities of Commuters

Ashish Dandekar*, Stéphane Bressan*, Talel Abdessalem†, Huayu Wu‡ and Wee Siong Ng‡
*School of Computing,
National University of Singapore, Singapore
Email: ashishdandekar@u.nus.edu, steph@u.nus.edu
†Télécom ParisTech, Paris-Saclay University, Paris, France
Email: talel.abdessalem@telecom-paristech.fr
‡Institute of Infocomm Research,
A*STAR, Singapore
Email: huwu@i2r.a-star.edu.sg, wsng@i2r.a-star.edu.sg

***Abstract*—Urban planning, development and management authorities and stakeholders need to understand and analyse the mobility patterns of urban dwellers in order to manage sociological, economic and environmental issues. Simulation is indispensable a tool for authorities and stakeholders to better design, operate and control the mobility infrastructures of smart cities. We propose an approach for the simulation of trajectories in communities of commuters. We identify communities of public transport commuters from historical automated fare collection card data using spatial latent Dirichlet allocation. We further aggregate the historical automated fare collection card data to create statistical models of visits and movements of commuters in each community. We use statistical models to simulate trajectories of synthetic individual commuters. We empirically evaluate how the synthetically generated trajectories are typical of their community of commuters and realistic.**

I. Introduction

Urban planning, development and management authorities and stakeholders need to understand and analyse the mobility patterns of urban dwellers in order to manage sociological, economic and environmental issues. Urban development authorities need to study mobilities of commuters so as to foresee mobility infrastructure requirements. Simulation serves as a faithful means to help in the study of mobility.

Human mobility in urban areas is the movement of an individual as much as the dynamics of communities moving from one place to the other. An urban area comprises of different residential areas from where people commute to work places on daily basis. A faithful simulation of human mobility in the urban setting should be able to not only capture spatiotemporal patterns of movements of individual commuters but also adhere to latent patterns of movement as a member of communities.

In this work, we propose an approach to simulate trajectories of commuters which are typical of their community structure. We apply spatial adoption of Latent Dirichlet Allocation to historic automated fare collection card data of a public transportation network to find latent communities of commuters. We use the learned community structure to synthesize the commuter transportation graph for each community. We further propose a mechanism which uses random walk on the graph along with latent topics learned from LDA to generate trajectories for a synthetic individual. Experiments show that simulated trajectories conform to the underlying hidden community structure.

The rest of the paper is organized as follows. Section II delineates related work. Section III presents the proposed mechanism to generate trajectories. We present experiments and evaluation in Section IV. We conclude the paper by discussing the work underway in Section V

II. Related Work

Related work spans three different domains of research, namely *Urban Computing*, *Latent Dirichlet Allocation (LDA)* and *Human Mobility*.

Urban Computing [1] is a process of acquisition, integration, and analysis of big and heterogeneous data generated by diverse sources in urban spaces, such as sensors, devices, vehicles, buildings, and humans, to tackle the major issues that cities face. It also helps to understand the latent trends and foresee the development of the city. Public transport data has been studied by authors for the betterment of the mobility of the citizens of the cities. In [2]–[4], London public transport has been widely studied for exploring traveling behaviors, minimizing traveling time and finding communities of citizens. Ferris et al. [5] have developed an app *OneBusAway* to reduce the waiting time of commuters by providing real-time bus arrival information in King county, Washington. To identify tourists from the daily commuters of the public transportation services Xue et al. [6] have devised a method and tested it on the data from Singapore. Montis et al. [7] have found communities of commuters to help in the demarcation of the sub-regions in Sardinia. In [8], authors have used to spatio-temporal data generated from social networks to find mobility patterns in an urban area.

Blei et al. [9] have proposed **Latent Dirichlet Allocation** (LDA) - a soft clustering technique used for finding latent topics by intuitively capturing the co-occurrence of the words in the textual corpora. Till the date, it is a widely used

TABLE I
NOTATIONS

Symbol	Description
$\mathcal{C}$	Set of commuters
$\mathcal{L}$	Set of locations
K	Total number of topics
$\mathcal{M}_c$	Mobility of a commuter c
θ_c	Topic distribution of a commuter c
ϕ_k	Location distribution of a topic k
$\mathcal{C}_k$	Community of commuters with topic k
G_k	Commuter Transportation Graph pertaining to community $\mathcal{C}_k$

technique for topic discovery. In [10], [11], authors have altered the original model to handle geospatial data. LDA has been used to find the group of places in cities using the check-in data from LBSN Foursquare users [12]–[14]. They have shown that LDA has the ability to cluster the places based on the hidden user interests than their geographical proximity. In [15], Ashish et. al. have adopted and extended LDA to work with spatiotemporal urban data.

Different datasets have been studied in literature to perform statistical analyses and hence design models to generate **Human Mobility** which adheres to the observations. González et. al. [16] have studied a large mobile phone dataset to find that travel distances and radius of gyration of individuals follow power-law behaviour. By studying entropy of individual movement, Song et. al. [17] ascertain that the human mobility is highly predictable. ORBIT [18] is a trajectory generation model which captures the heterogeneity in the human mobility. Trajectories generated using SLAW [19] conform to the power-law observations made by González et. al.. SWIM [20] and EPR [21] take into account location preference of individuals while generating trajectories.

In the current work, we exploit the duality of LDA to cluster commuters and locations in order to simulate trajectories of humans in urban areas.

III. METHODOLOGY

We propose an approach to simulating trajectories of commuters which are typical to the community structure they belong to. We begin with explaining the method we employ to find latent communities of commuters using logs of their spatiotemporal movements. It is followed by the commuter transportation graph synthesis and procedure to generate trajectories for a synthetic individual. Before we explain the method, we introduce the notation which is used throughout the paper.

Let $\mathcal{C}$ denotes set of commuters and $\mathcal{L}$ denotes set of locations. The list of locations which a commuter visits in the given time constitutes her *mobility*. The mobility of a commuter c, $\mathcal{M}_c$ is a multiset $\{l|l \in \mathcal{L}\}$. Each element of $\mathcal{M}_c$ is called *visit* of a commuter. Table I enlists all notations which are used in the current work.

A. Communities of Commuters

Latent Dirichlet Allocation [9] (LDA) is widely used in the domain of Natural Language Processing to find topics in the collection of documents. LDA assumes each document as a bag of words. Each topic i found by LDA is represented as a probability distribution (ϕ_i) over words whereas each document d is represented as a probability distribution (θ_d) over K topics, K being an input to the model. LDA is an probabilistic generative model, which one can use to generate new documents after learning parameters using training data.

We adopt spatial LDA [15] to find latent topics in the geospatial data. We treat mobilities of all commuters as a collection of documents and locations as the vocabulary for LDA. More precisely, for a commuter c, her mobility $\mathcal{M}_c$ stands as a document and different locations visited by her as words in that document. As explained earlier, LDA is a probabilistic clustering technique which assigns each document to every single topic with a certain probability. We want to cluster commuters into the clusters pertaining to typical mobility patterns. We, therefore, apply a threshold to cluster them in disjoint communities.

After running spatial LDA on the corpus of mobilities, we learn θ_c - topic distribution of each commuter c and ϕ_k - location distribution of each topic k. Let $\mathcal{C}_k = \{c|c \in \mathcal{C}, \theta_c^k \geq \tau\}$ denote a community of commuters with topic k, for some threshold $\tau \in [0, 1]$. It can be observed that $\mathcal{C} = \cup_{i=1}^{K} \mathcal{C}_i$. So, the communities of commuters partition the set of commuters $\mathcal{C}$ in K partitions.

B. Commuter Transportation Graph Synthesis

Using the partitions of commuters generated by LDA, for each topic, we generate a weighted directed graph of locations using the mobilities of those users. For a given topic k, the graph comprises of locations as vertices and an edge is added between two vertices if a commuter from the partition $\mathcal{C}_k$ has commuted between these locations.

Formally, for a given topic k, we create a weighted directed graph $G_k(V_k, E_k)$, called as Commuter Transportation Graph, where $V_k = \mathcal{L}$ and $E_k = \{(v_1, v_2) \mid v_1, v_2 \in V, \exists\ c \in \mathcal{C}_k\ v_1, v_2 \in \mathcal{M}_c\}$. Each edge is weighted by number of commutations observed between two locations in the community $\mathcal{C}_k$.

C. Trajectory Generation

We perform a random walk on the graph to generate the trajectory of a synthetic user. For a given topic k, so as to pick a starting point of the trajectory, we non-uniformly sample a location from the location distribution ϕ_k for topic k. Next location is chosen by following one of the outward edges of the previous node in the commuter transportation graph depending on the weight. If a node does not have any outgoing edge then the next location is again sampled from ϕ_k. We repeat the same procedure until we generate the desired number of visits for a synthetic individual.

The procedure to generate n visits for a synthetic commuter in a community z is given in Algorithm 1.

Algorithm 1 Algorithm for generating trajectories typical to a given community

Input: Community z, ϕ_z, G_z
Output: Trajectory l

$l \leftarrow []$
$l_0 \sim Multinomial(\phi_z)$
for $i = 1$ to $n - 1$ **do**
  if $(outdegree(l_{(i-1)} \in V_z) = 0)$ **then**
    $l_i \sim Multinomial(\phi_z)$
  else
    Choose l_i as one of the out-neighbors of $l_{(i-1)}$ with probability proportional to weight of the edge $(l_i, l_{(i-1)})$ in G_z
  end if
end for
return l

Location probability distributions, ϕ_is, essentially encode the community biases towards different locations. Sampling from ϕ_i ensures that generated trajectories adhere to communities of commuters.

IV. Results

In this section we evaluate the effectiveness of the proposed method of trajectory generation. We learn the community structure of commuters, which we further use to generate trajectories, of public transportation network of a city. We propose a metric to check how typical the generated trajectories are to underlying communities.

All programs are run on a Linux machine with quad core 2.40GHz Intel® Core i7™processor and 8GB memory. Python® 2.7.6 is used as a scripting language. We have used the Java library, *jLDADMM*, developed by Dat et al. [22] for finding the latent topics using LDA[1] and modified the source to adapt to proposed extension of LDA. For the graph computation, we use Python wrapper for an highly-efficient *igraph*[2] library written in C.

A. Dataset Description

The dataset comprises of tappings registered on automated fare collection cards of the public transportation system of a city which consists of metro stations and bus stops. The dataset contains tappings of around 3.5 million commuters over a period of one month. Simple frequency based analysis shows that the majority of the commuters follow a characteristic pattern during weekdays *viz.* commuting to workplace in the morning and commuting back to home in the evening. We filtered data for a typical weekday with 40,000 regular commuters.

B. Typicality

Metric: As explained earlier, location probability distributions, ϕ_is, capture the community structure of commuters. We use them to predict the community of trajectories of synthetically generated individuals. We vectorize a generated trajectory as $|\mathcal{L}|$-dimensional vector, say $\bar{t}$. Inner product of $\bar{t}$ with ϕ_i denotes the probability of the trajectory to be in community i. So the community for trajectory $\bar{t}$ is predicted as follows:

$$k_{\bar{t}} = \underset{i \in [1..K]}{\operatorname{argmax}} \phi_i \cdot \bar{t}$$

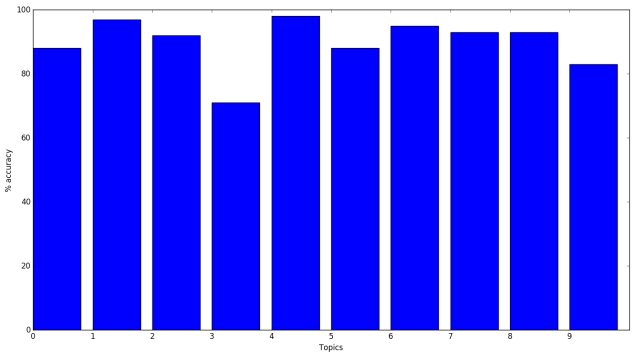

Fig. 1. Typicality evaluation

Results: We perform 2000 Gibbs sampling iterations for spatial adoption of SLDA to find the latent topics. We set threshold τ as 0.4 and find communities of commuters. We use these communities to construct the commuter transportation graph.

For each community we generate 1000 trajectories each of length 10. Then we calculate accuracy of the proposed method for each community as the fraction of trajectories out of 1000 that are correctly predicted to be from the same community. Figure 1 shows the result of the experiment. It can be observed that the proposed method generate trajectories which are typical to the communities of commuters.

V. Conclusion

In this paper, we propose a method to generate trajectories which are typical to the community of commuters in urban areas. We conduct experiments on a real-world dataset of the public transportation network of a city. We empirically show that the proposed method does generate trajectories which adhere to the community structure.

The proposed method is able to capture mobility patterns of the community rather than mobility patterns of a synthetic individual. We are working on how generated trajectories can be made more realistic, in the sense that they capture mobility patterns of an individual as much as mobility patterns on communities.

Acknowledgment

This research is funded by research grant R-252-000-622-114 by Singapore Ministry of Education Academic Research Fund (project 251RES1607 - Janus: Effective, Efficient and Fair Algorithms for Spatio-temporal Crowdsourcing) and is a collaboration between the National University of Singapore,

[1] http://jldadmm.sourceforge.net/
[2] http://igraph.org/redirect.html

Télécom ParisTech and Singapore Agency for Science, Technology and Research.

REFERENCES

[1] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 38, 2014.

[2] N. Lathia and L. Capra, "Mining mobility data to minimise travellers' spending on public transport," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1181–1189.

[3] ——, "How smart is your smartcard?: measuring travel behaviours, perceptions, and incentives," in *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 2011, pp. 291–300.

[4] N. Lathia, D. Quercia, and J. Crowcroft, "The hidden image of the city: sensing community well-being from urban mobility," in *Pervasive computing*. Springer, 2012, pp. 91–98.

[5] B. Ferris, K. Watkins, and A. Borning, "Onebusaway: results from providing real-time arrival information for public transit," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 1807–1816.

[6] M. Xue, H. Wu, W. Chen, W. S. Ng, and G. H. Goh, "Identifying tourists from public transport commuters," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1779–1788.

[7] A. De Montis, S. Caschili, and A. Chessa, "Commuter networks and community detection: a method for planning sub regional areas," *The European Physical Journal Special Topics*, pp. 75–91, 2013.

[8] M. Al-Ghossein and T. Abdessalem, "Somap: Dynamic clustering and ranking of geotagged posts," in *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016, pp. 151–154.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, pp. 993–1022, 2003.

[10] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 186–194.

[11] B. Hu, M. Jamali, and M. Ester, "Spatio-temporal topic modeling in mobile social media for location recommendation," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 1073–1078.

[12] X. Long, L. Jin, and J. Joshi, "Exploring trajectory-driven local geographic topics in foursquare," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 927–934.

[13] K. Joseph, C. H. Tan, and K. M. Carley, "Beyond local, categories and friends: clustering foursquare users with latent topics," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 919–926.

[14] Y.-S. Cho, G. Ver Steeg, and A. Galstyan, "Socially relevant venue clustering from check-in data," in *11th Workshop on Mining and Learning with Graphs, MLG–2013*, 2013.

[15] A. Dandekar, S. Bressan, T. Abdessalem, H. Wu, and W. S. Ng, "Detecting communities of commuters: graph based techniques versus generative models," in *24th International Conference on Cooperative Information Systems (COOPIS 2016)*, 2016.

[16] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[17] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[18] J. Ghosh, S. J. Philip, and C. Qiao, "Sociological orbit aware location approximation and routing in manet," in *2nd International Conference on Broadband Networks, 2005.* IEEE, 2005, pp. 641–650.

[19] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "Slaw: A new mobility model for human walks," in *INFOCOM 2009, IEEE*. IEEE, 2009, pp. 855–863.

[20] S. Kosta, A. Mei, and J. Stefa, "Small world in motion (swim): Modeling communities in ad-hoc mobile networking," in *2010 7th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*. IEEE, 2010, pp. 1–9.

[21] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010.

[22] D. Q. Nguyen, "jLDADMM: A Java package for the LDA and DMM topic models," http://jldadmm.sourceforge.net/, 2015.