Estimation of Human Mobility Patterns and Attributes Analyzing Anonymized Mobile Phone CDR:

Developing Real-time Census from Crowds of Greater Dhaka

Ayumi Arai[1] and Ryosuke Shibasaki[1,2]

[1] Department of Frontier Science, The University of Tokyo
[2] Center for Spatial Information Science, The University of Tokyo
{arai,shiba}@csis.u-tokyo.ac.jp

Keywords. CDR, human mobility, demographic attributes, machine learning

Abstract. Anonymized mobile phone CDR allows us to capture dynamics of mass population movement where individual trajectories are still traceable. While, outcomes of research analyzing CDR merely show distribution of people or crowds, which are aggregation of mass trajectories without any attributes. To further investigate hidden properties of human mobility in CDR, it is critical to analyze such data in combination with secondary datasets. This project develops Real-time Census of Greater Dhaka from CDR. It represents population composition of Greater Dhaka and is labeled with demographic attributes such as sex, age groups, and occupational types. Algorithms developed in this project can be applicable to CDR in other places wherever census is available.

1 Background

Emergence of large-scale datasets such as GPS logs and Call Detail Records (CDR) of mobile phone has advanced understanding of human mobility. Anonymized CDR allows us to capture dynamics of mass population movement where individual trajectories are still traceable. Recent explosion of research on human mobility is divided broadly into two areas based on the method of analyses [1]. One area is developed by quantitative approaches to model properties of human mobility. Song *et al.* models decreasing likelihood to explore a new place to visit in a long term, which follows power law decay [2]. It indicates that people tend to visit highly frequented locations, which are home and work locations, repeatedly [3]. The other area has evolved with the development of data mining techniques to learn frequent patterns and association rules of human behavior from large and complex datasets. Isaacman *et al.* proposes an algorithm to identify home and work locations of mobile phone users applying clustering and regression technique to CDR [4]. Li *et al*. mines similarities among GPS device users based on sequence properties of people's trajectories and hierarchy properties of location histories [5]. The study considers people who have similar location histories would share similar interests and preferences. While the advantage in capturing the trajectories of human mobility, outcomes of research analyzing CDR merely

shows distribution of people or crowds, that is, aggregation of mass trajectories without any attributes. It is because CDR is anonymized, which mitigates privacy concerns and at the same time allows tracing individual trajectories. To further investigate hidden properties of human mobility, it is critical to analyze such data in combination with secondary datasets [6].

Besides the data obtained through ubiquitous means, conventional survey data has long been contributing to studies on human activity-travel patterns primarily for urban planning and transportation. Data is collected through a survey, which collects basic demographic attributes, means of transportation, and origins, destinations, and purposes of movement with time stamps. Due to limitations in capturing human mobility through such an interview survey, studies tend to focus on correlations between activity-travel patterns and demographic attributes, such as employment status, gender, and presence of children [7], and job types [8]. Although the activity-travel pattern captured in the study area is just shown as descriptive one, it is significant that their research findings can link demographic attributes with human mobility patterns to some extent. It indicates anonymized CDR can be linked with such survey data through mobility patterns as a common key property.

2 Purpose

This study aims to develop a system to create Real-time Census of Greater Dhaka[1], which is the visualization of mass population trajectories. It represents population composition of Greater Dhaka and is labeled with demographic attributes. An advantage of Real-time Census is labeling, which allows filtering specific population groups according to the purpose of application. For instance, it can be used to address the containment of infectious diseases by filtering the movement of higher-risk population such as males at a certain age group. Data input to operate the system is CDR and census, both of which exist globally. CDR is routinely collected data by the cellular network provider for optimizing their network and billing purposes; thereby, the system is applicable wherever mobile networks are available. Census has been conducted in more than 200 countries as an important baseline survey to address global issues with the promotion of United Nations [9]. The system, which does not require many parameters, is expected to expand users of the system and accelerate the use of CDR.

[1] Greater Dhaka includes parts of Dhaka District in Bangladesh; Dhaka city cooperation and surrounding Thanas, Savar Upazila, and Karaniganj Upazila. It also covers parts of Narayanganj District, Gazipur District, and Narsingdi District.

3 Data

3.1 Designing architecture of the system

To design base algorithms for the system, anonymized CDR of seven million people in Greater Dhaka and Person Trip survey data (PT data) are analyzed. CDR is provided by Grameenphone that is a telecommunication operator and has the largest mobile phone customer base in Bangladesh. It includes the record of time and tower location when people dial during six months between 1st August 2013 and 31st January 2014. PT survey is an interview-based Origin-Destination survey conducted by Japan International Cooperation Agency in 2009. It includes demographic attributes and one-day travel-activity records of 75,000 people, residing in Greater Dhaka. In this project PT data is used to estimate demographic attributes of anonymized CDR where the mobility pattern is as a common key variable among CDR and PT data.

3.2 Addressing biases of CDR

Additional two datasets are used to address two types of biases of CDR; one is deriving from mobile phone user behavior and the other is sampling biases. Mobile phone user behavior causes biases because the timing of recording CDR associates with that of mobile phone usages [2]. Time, which is recorded as part of CDR, dose not necessarily indicate time of departure or arrival when a series of different locations and time are recorded. However, there is very limited data that can link individual activity patterns and calling behavior. Thus, the University of Tokyo conducted a household survey, Survey on Patterns and Activities for Comprehensive Exploration of Mobile Phone Users in Dhaka (SPACE). It samples 810 households and interviews all household members on mobile phone ownership and patterns of mobile phone usage. SPACE is designed to collect demographic attributes and daily travel-activity patterns as well as to be used for validation to test algorithms developed through this project.
Sampling bias is one of typical problems for research, which analyzes data acquired through specific devices [1]. Because the data excludes a certain group of population who do not use the device, it causes serious problems particularly when analysis results are applied to address issues in society. To address the bias, data from Household Income and Expenditure Survey 2011 (HIES) is compared with CDR. HIES is a nation-wide household survey conducted by Bangladesh Bureau of Statistics in 2011, which samples a 0.52% of households of Greater Dhaka. It is adjusted to represent population groups of sampled areas and includes basic demographic attributes such as sex, age, and job types.

4 Methodology

4.1 Estimation of home and work locations

Algorithms of developing Real-time Census are designed by a four-step approach. First, important places, such as home and work locations, for seven million people are

estimated by analyzing CDR. Following the method developed by Isaacman *et al.* [4] identifies important locations based on the amount of time spending and frequencies of visit of locations. Then, core hours for work, which are between 1PM and 5PM on weekday, and core hours at home, which are between 7 PM and 7AM, are taken into account to distinguish home and work locations.
Once home and work locations are estimated, paths of individual movement are visualized as trajectories. As the method to generate the path, which connects discrete time and location data of mobile objects, various interpolation methods have already been developed. However, they virtually focus on mathematical approaches for spatial interpolation and yet consider activities or events associated with the timing of record [10]. Disregarding the timing of record in CDR, at which people use mobile phone, could lead to biases because the timing is not random but associated with certain types of locations [11] and activities [12]. To reduce the bias, probabilities of using a mobile phone are calculated using data from SPACE. The probabilities are likelihood of making phone calls, which is associated with specific types of locations and activities. It provides probability functions of the timing to start traveling between two different activities and locations.

4.2 Estimation of demographic attributes of CDR

Support Vector Machines (SVM).
Then, demographic attributes are estimated using PT data as training data and CDR as test data employing SVM, supervised learning. Travel-activity patterns are sorted out and clustered from PT data based on several demographic attributes, including sex, age groups, and occupational types. Both travel-activity patterns from PT data and CDR are transformed into hourly basis activity for training and testing procedures. Occupational types are used as a primary label for training. It is because several aspects, closely related to occupational types, are inferred as important factors, affecting human mobility. For instance, Mo *et al*. indicates chances of being at home at night as well as time spending at work place during weekend are useful features to estimate job types [13]. Szell *et al*. argues human behavior inevitably follows some forms of patterns due to social ties such as sleeping at home and working at the office [14], which tend to occur at specific places.

Non-negative Matrix Factorization (NMF).
Along with SVM, demographic attributes of CDR are separately estimated employing non-negative matrix factorization (NMF), semi-supervised machine learning. NMF is additionally chosen since this study aims to design algorithms, which can be implement with insufficient training data. Part of CDR is labeled for constructing training data using 925 samples from SPACE whose home and work locations, and mobile phone usage patterns are similar to part of individuals in CDR. Then, remainder of the CDR is used as test data for labeling. NMF identifies principal components by decomposing data into low-dimensional characterizations [15]. In this study it learns to represent daily activity schedule of a week as a linear combination of basis schedules,

which are localized feature of the activity schedule. Dataset for the activity schedule is constructed from CDR based on time spent at home and work locations where individual activity schedule is regarded as a 7×24 matrix V. Each column of the matrix contains 24 non-negative values, which represent activity schedules of a day: being at home, working, or else. Then, it constructs approximate factorizations of the form as such:

$$V \approx WH \tag{1}$$

where columns of W are basis schedules. Each column of H is an encoding, which is one-to-one correspondence with a basis schedule in V. Part of labels for the activity schedule are generated with a conditioned Hidden Markov Model since it is originally discrete.

4.3 Adjusting sampling biases

Third, biases deriving from sampling CDR are adjusted by comparing labeled CDR with HIES. The datasets can be linked using home locations as a key variable where HIES is sampled based on residence locations. Home location of CDR is already estimated in the first step. For the adjustment, mobility patterns of small children and elderly people, who seldom use mobile phones, need to be taken into account. As they are not included in CDR, PT data and SPACE, which captures principal activities of such population groups, are used as supplement.

4.4 Validation

Last, validation of the algorithm is conducted using part of SPACE as grand truth. In this step, estimation accuracy of SVM and NMF is tested.

5 Expected outcomes

This project develops algorithms of developing Real-time Census where demographic attributes of anonymized mobile phone users are estimated. The algorithm transfers CDR, which is primarily used to understand mass population movement and distribution, into the aggregation of trajectories with labels of sex, age groups, and occupational types. The labeled trajectories are color-coded by attributes.

The algorithms can be applicable to CDR in other areas wherever census is available. Real-time Census is useful for various applications. It enables us to understand what kind of people exists in which place at which time. For instance, it can improve disaster preparedness by estimating the location of vulnerable people against disasters at the occurrence of disastrous events. Overlaying Real-time Census with secondary information such as land use and road networks, it can also contribute to further development of city planning and efficient transportation.

6 Acknowledgements

CDR is provided by Grameenphone under the MoU between The University of Tokyo. I am grateful to AGILE for providing opportunities to develop this project through the 2nd AGILE PHD School 2013. I would like to thank participants of the PHD School for their valuable comments.

7 References

1. Wang, D., D. Pedreschi, C. Song, F. Giannotti, & A. L. Barabasi, 2011, August. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1100-1108. ACM.
2. Song, C., Koren, T. Koren, P. Wang, & A. L. Barabási. 2010. Modelling the scaling properties of human mobility. *Nature Physics*, *6*(10), 818-823.
3. González, M. C., C. A. Hidalgo, & A. L. Barabási. 2008. Understanding individual human mobility patterns. *Nature*, 453(7196), 779-782.
4. Isaacman, S., R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, & A. Varshavsky. 2011. Identifying important places in people's lives from cellular network data. In *Pervasive Computing*, 133-151. Springer Berlin Heidelberg.
5. Li, Q., Y. Zheng, X. Xie, Y. Chen, W. Liu, & W. Y. Ma. 2008. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. 298-307. ACM.
6. Lu, X., L. Bengtsson, & P. Holme. 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences*, *109*(29), 11576-11581.
7. Pas, E. I. 1984. The effect of selected sociodemographic characteristics on daily travel-activity behavior. *Environment and Planning A*, 16(5), 571-581.
8. Kitamura, R., C. Chen, R. M. Pendyala, & R. Narayanan. 2000. Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation*, *27*(1), 25-51.
9. United Nations. 2007. Principles and recommendations for population and housing censuses revision 2. Retrieved 15 December 2013 from http://unstats.un.org/unsd/demographic/sources/census/docs/P&R_%20Rev2.pdf
10. Miller, H. J. 2005. A measurement theory for time geography. *Geographical analysis*, *37*(1), 17-45.
11. Sohn, T., Li, K. A., Lee, G., Smith, I., Scott, J., & Griswold, W. G. 2005. *Place-its: A study of location-based reminders on mobile phones*. 232-250. Springer Berlin Heidelberg.
12. Becker, R. A., R. Cáceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, & C. Volinsky. 2011. A tale of one city: Using cellular network data for urban planning. *Pervasive Computing, IEEE*, *10*(4), 18-26.
13. Mo, K., B. Tan, E. Zhong, & Q. Yang. 2012. Report of task 3: Your phone understands you. Retrieved 15 December 2013 from https://research.nokia.com/files/public/mdc-final131-mo.pdf
14. Szell, M., R. Sinatra, G. Petri, G., S. Thurner, & V. Latora. 2012. Understanding mobility in a social petri dish. *Scientific reports*, *2*.
15. Eagle, N., & A. S. Pentland. 2009. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7), 1057-1066.