An Interactive Analytics Tool for Understanding Location Semantics and Mobility of Users Using Mobile Network Data

Manoranjan Dash, Gim Guan Chua, Hai-Long Nguyen, Ghim-Eng Yap, Cao Hong, Xiaoli Li, Shonali Priyadarsini Krishnaswamy
Institute for Infocomm Research
*A*Star, Singapore 138632*
{*dashm, ggchua, nguyenhl, geyap, hcao, xlli, spkrishna*}
@i2r.a-star.edu.sg

James Decraene, Amy Shi-Nash
R&D Labs, Living Analytics, Group Digital Life,
31b Exeter Road, Comcentre 2
Singapore Telecommunication Limited, Singapore
{*jdecraene, amyshinash*}*@singtel.com*

***Abstract*—Knowledge about population distribution of planning areas helps in making urban development decisions. Two important criteria are: "where do people live?" and "where do they work?" In this paper we propose methods to find home and workplaces from mobile network data. Home and work places are essential for discovery of mobility profiles of users. Validation of home and workplace prediction is not straight forward. We validate our methods using correlation with external data. Validation results show that even though a single cellular provider has only a portion of the entire population as its users, distribution of home and work places predicted using its mobile network data match that of government statistics. On the basis of this matching, we can have faith in distributions of more difficult statistics extracted from mobile network data which are difficult to obtain from external sources. We implemented an interactive system to show various distributions such as people living and working in different planning areas, and people working in different job sectors such as manufacturing. Interesting relationships are found by calculating joint distributions, e.g., where do people, living in a planning area, work, and vice versa. Planning areas are ranked by the average distance travelled from home to work. Another interesting fact we extract is *balance*. Balance of a planning area is high if people live and work there; it is low if people living in a planning area work in other planning areas. We extend these statistics to regions which consist of many planning areas. The goal of this interactive system is to understand location semantics and mobility of users to aid in making urban development decisions. A video recording with subtitles is uploaded in http://www.youtube.com/watch?v=mo-7-DsCymw.**

I. Introduction

Analysis of mobile network data can find knowledge about the mobility profiles of people of a city. Such knowledge can be used in many applications such as city wide sensing [4], product advertisement [2], pollution exposure, route marking and tracking [5], early warning systems [6], traffic management, social networking and community finding [7]. A constant feature of any mobility profile is the knowledge of home and work places. For example, a mobility pattern or motif can be like "9:30am: left home, 9:57am did some errand, 10:55am: some outdoor activities, 11:52am: stayed for some time in another person's home, etc" [3]. Accurate determination of such motifs require accurate prediction of home and work-place. In this work, we present a demo based on methods for predicting and validating home and work places.

According to some industry research estimates, only about 10% of phones manufactured in a year have the GPS capability; our focus in this paper concerns the remaining 90% users (GSM – network based) [1]. GSM network-data lack in accuracy of positioning compared to GPS data; however, it is the most popular means of mobile communication. A goal of this research is to find out whether the large volume of GSM data can compensate for such loss of accuracy.

For this research project, three months of mobile network data is used. Mobile network data is the service log when a mobile phone is connected to the mobile network. Each record or event or transaction contains an anonymised ID, latitude and longitude of the cell tower, time stamp of the event and service type. Anonymised IDs are machine generated via a two-step non-reversible AES encryption and hash process. This means it is not possible to trace back the original IDs. There is no personal information about mobile subscribers in the data set, nor any content of calls or SMSs. Lat-lon of a cell tower covers a range of 50 to 200 meters. The insight in the visualisation tool is aggregated at the planning area level.

For home prediction, we extract a new feature called *"inactivity"*. If a phone is idle for more than a period of time (say, 5 hours) excluding the automated location update events, we increase the inactivity count for the corresponding cell tower. Moreover, to improve the accuracy of prediction, we further extract useful information from external web mapping services (such as www.StreetDirectory.com) that tells whether the predicted location (latitude-longitude) is a residential/non-residential place.

To predict the workplace, we exploit two observations

[1] http://searchengineland.com/cell-phone-triangulation-accuracy-is-all-over-the-map-14790

of human behavior: (1) people go to work on most of the weekdays and rarely on the weekends, and (2) they usually stay at their workplaces during typical office hours (2 pm to 5pm). We choose this time range since it covers for many types of jobs, including office jobs as well as shift jobs.

Validation of home and workplace prediction is not straight forward [1]. Third party information is used to validate. This third party information can be based on a survey study in which participants agree to disclose their mobile network data. The number of participants in such survey and their distribution holds the key to reliability of the validation. Another third party information can be based on various demographic statistics available from different sources. This approach is indirect in that summary statistics from mobile phone data is compared with those from the third party. In this work, we validate the predicted results using both, i.e., the direct (survey) and indirect approaches. Reliability of the validation is enhanced by cross validating using different demographic statistics. Validation of home prediction is easier than validation of workplace prediction. Rarely any external source provides statistics about distribution of population based on workplace.

Validation results show that predicted distribution statistics correlate strongly to the actual distribution statistics. A Pearson correlation coefficient of 0.95 was achieved for home prediction. This result further strengthens the motivation to extract useful information from mobile network data which is not easy to extract from external sources. For example, one can rarely find information about distribution of people working in different planning areas. But, using good heuristics, such statistics can be found from mobile network data, and we can be more or less confident about it.

Using these distribution statistics, we can extract information to aid in making urban development decisions. For example, planning areas were ranked by their average distance travelled from home to work. Planning areas were compared by their balance. Balance of a planning area P is defined as the ratio of people working in P who also live in P and total working population among people living in P.

In this paper, we illustrate an implementation of our home-workplace prediction method. An interactive analytic tool for understanding home and workplace semantics is created with color gradient and 3D features including zooming, rotating, panning, and flowing arrows.

II. Methodologies and Results

In this section we first describe the methodologies used for prediction, and then briefly describe the results.

Overview of the entire process

1. Preprocess the mobile network data
2. For each anonymized ID
   - Extract features
   - Based on the extracted features, predict the home and workplace locations (latitude and longitude of the corresponding cell tower)
3. Using third party source, validate home and workplace

Preprocessing: Since mobile network data is huge and dirty, several preprocessing methods are performed to clean it, such as noise removal, oscillation removal, and in-transit removal. Transactional data in transit is removed by using two thresholds – velocity threshold and distance threshold. In-transit mobile network data are not necessary for determining home and workplaces.

Method to Predict Home: We have devised a novel method to predict home using inactivity. Inactivity is defined as *no activity for more than threshold time except for location update event*. In our experiments, we used five hours as the *inactivityThreshold* in order to model the sleeping hours. This works well for shift workers as well. A location update event is an automated event initiated by the cellular provider to maintain connectivity whilst users are moving. For each anonymized ID, we compute the total number of inactivities for each tower. The tower with the highest number of inactivity is predicted to be the home location. Third party data (www.streetdirectory.com) is used to improve the accuracy. If the predicted home location is not in a residential place, the residential location with the highest inactivity is predicted as the home.

Method to Predict Workplace: The proposed method is based on the fact that people go to work on weekdays regularly and rarely on weekends. This ratio is multiplied by duration during weekday from 2pm to 5pm. By doing so we give more importance to towers which have high duration during 2pm to 5pm on weekdays. We choose the time range from 2pm to 5pm because this time range is the most probable range for people to be present at their work place. This works well even for shift workers (at least for two shifts).

Validation: Two types of validations are performed. (1) Direct approach – validate against a survey of 4515 panelists, who disclosed their home location. (2) Indirect approach – various demographic statistics, such as government data, are used. We first group people according to their planning areas. The probability distribution for the whole city is then aggregated. We compare this statistics with available demographic statistics. Correlation between two distributions is used as a measure for validation.

Results: In this study we use mobile network data of 3,875,254 anonymized users of Singapore. There are around eight billion transactions over three months (May - July, 2013). We compared the proposed method with the method in [1]. We set the best possible thresholds for [1]. Correlation is calculated by comparing the predicted distribution with the statistical demographics information from Urban Redevelop-

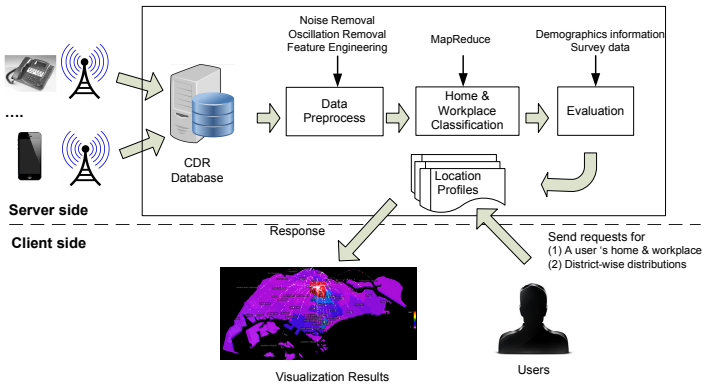

Figure 1. System Architecture

ment Authority, Singapore (URA)[2]. Correlation coefficient is 0.93 for the proposed method without combining with 3rd party data, and 0.95 when we combine. For Ahas-etal-2010 method, the results are 0.92 and 0.93 respectively. Validation of work place is not easy. Except for manufacturing, planning area-wise distribution for other job sectors are not available. For manufacturing sector, correlation between predicted distribution and 3rd party statistics is 0.92.

Sample Data with Ground Truth: A sample of 4515 anonymized users agreed to disclose their home locations. Our predictions were correct (within 3km of declared home location) for 88% users.

III. System Design and Demonstration

The system is designed as a client server application. Figure 1 shows an overview of the system architecture. The client is a web application written using HTML5, WebGL and Json. The client can visualize various demographic distributions for home and workplaces.

The server collects call transaction data from people, builds up mobile network data, extracts knowledge from those databases, and processes user queries. Since raw mobile network data is dirty to some extent, it is necessary to clean the data, such as, noise removal, oscillation removal, and feature engineering. After cleaning, the data is still too huge to be processed on a single machine. MapReduce techniques are deployed to process the data in parallel on a cluster of computers.

Client-side visualization program has the following features:

1) It uses HTML5 + WebGL to visualize distribution of data. Firefox browser is recommended as it enables loading of local files by default.
2) Three data sets are used: (1) planning area polygons, (2) home-workplace distribution data, home-manufacturing distribution data and balance data for planning areas, and (3) home-workplace distribution data and balance data for regions. Distance between home and workplace is calculated using these data sets. There are five regions and 55 planning areas in Singapore. Regions are east, west, north, north-east, and central. Each region consists of multiple planning areas. Data sets are all in CSV format. D3 (http://d3js.org/) is used to load and parse CSV files into data structures that can be more easily visualized.
3) For visualization, JavaScript 3D library three.js (http://threejs.org/) is used. The 3D scene consists of a base map of Singapore with planning area polygons superimposed over it. Interaction with the scene is via the mouse. Moving the mouse while pressing the left button will rotate the scene. The scroll wheel will zoom in and out based on the direction of scrolling. By moving the mouse while pressing the right button will pan around the scene.
4) Planning area polygons have been extruded to give it a 3D shape so that its height may be changed based on distribution values. Planning areas are colored using HSL model to reflect on magnitude of the distribution values. Color gradient goes from red for maximum to magenta for minimum. Visualizations of regions are done the same way as planning areas.
5) We have used a combination of curved line and moving particles to simulate the effect of the flow of distribution. Speed of the particles as well as their sizes corresponds to the distribution values.
6) The GUI consists of radio buttons and dropdown lists. It has been split into two: planning areas and regions of Singapore.
7) A top-5 and a bottom-5 buttons identify the top five and bottom five planning areas for each distribution.

Home-Work Distribution: A user can either choose a planning area for home to see distribution of workplaces of people living in that planning area, or choose a planning area for work to see distribution of home of people working in that planning area. Figure 2 shows workplace distribution of people living in Ang Mo Kio planning area.

Home-Manufacturing Distribution: Among all job sectors, only manufacturing sector can be validated. A user can either choose a planning area for home to see distribution of locations of manufacturing for people living in that planning area, or choose a planning area for manufacturing sector to see distribution of home of people working in manufacturing sector in that planning area.

Balance Distribution: As a by-product of visualization, we can determine how balanced a planning area is. A more 'balanced' planning area will be able to absorb its own population working in that planning area itself. Well balanced planning areas are: Changi, Tampines, Queenstown, Downtown Core, Yishun, etc. This result is supported by

[2]URA (www.ura.gov.sg) provides planning area-wise distribution of population in Singapore.

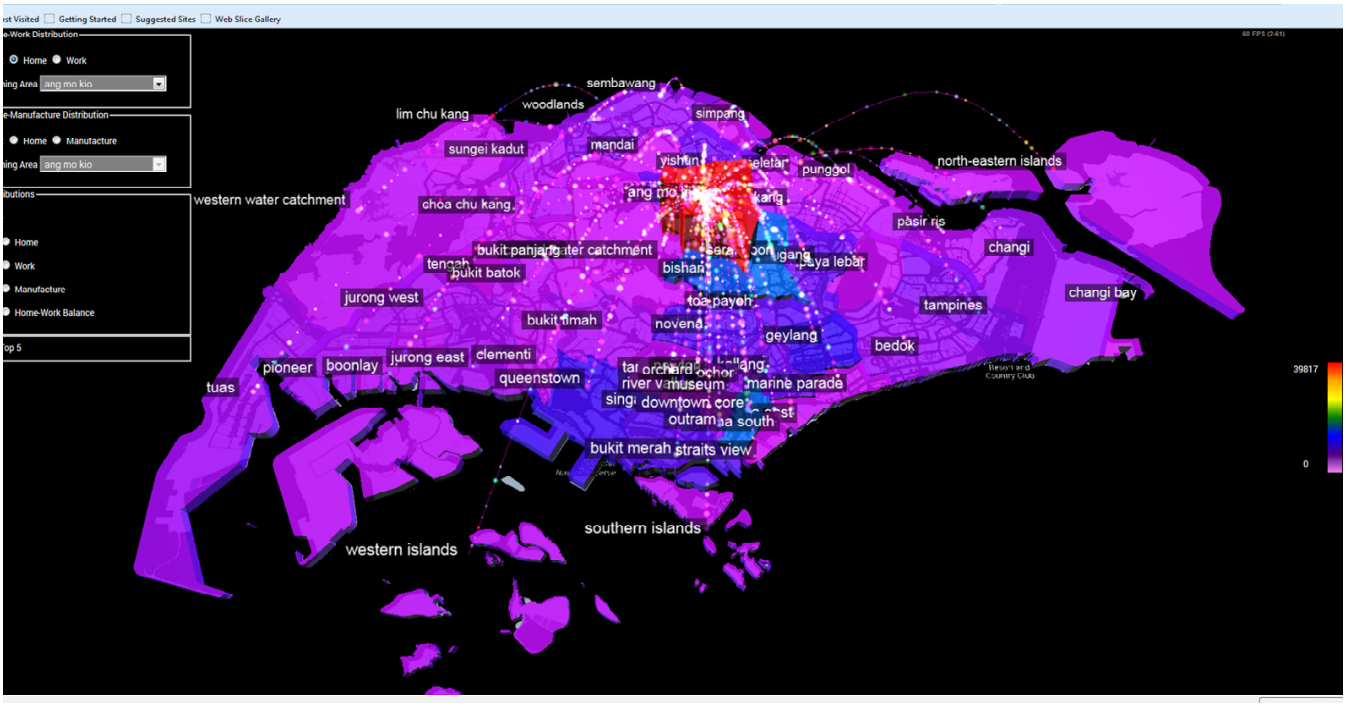

Figure 2. Home-Work Distribution: Home (Ang Mo Kio)

the **World Habitat Award** by united nations [3]. According to this study Jurong Island, which is largely a manufacturing and trading planning area, has low balance. Among the regions, central region, because of downtown area, has the highest balance, and the next highest is eastern region which includes Tampines planning area. North-east region, which has upcoming planning areas like Sengkang, has the lowest balance. Modern urban planning strongly favors high population density with development, which translates to high balance.

Travel Distance: A urban development planner will like to know distribution of distance travelled from home to work place. Using our system, the top five planning areas with the highest average trips from home to work are (in descending order): North-Eastern Islands, Lim Chu Kang, Simpang, Sungei Kadut and Western Water Catchment. The planning areas with the lowest average trips are (in ascending order): River Valley, Newton, Rochor, Museum, and Downtown Core. These results are very reasonable. For example, people of downtown core travel less distance to work, but people living in places like north-eastern islands have to travel a long distance to work.

A video recording with subtitles is uploaded in http://www.youtube.com/watch?v=mo-7-DsCymw.

[3]http://en.wikipedia.org/wiki/Tampines and http://www.hdb.gov.sg/fi10/fi10320p.nsf/w/AboutUsTown Tampines

REFERENCES

[1] R. Ahas, S. Silm, O. Jrv, E. Saluveer, and M. Tiru. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, 17(1):3–27, 2010.

[2] M. Couceiro, D. Suarez, D. Manzano, and L. Lafuente. Data stream processing on real-time mobile advertisement: Ericsson research approach. In *Mobile Data Management (MDM), 2011 12th IEEE International Conference on*, volume 1, pages 313–320. IEEE.

[3] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M. C. Gonzalez. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. ACM.

[4] S. S. Kanhere. *Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces*, pages 19–26. Springer, 2013.

[5] J. Lee, R. M. A. Mateo, B. D. Gerardo, and S.-H. Go. *Location-aware Agent using Data mining for the Distributed Location-based Services*, pages 867–876. Springer, 2006.

[6] C. Yang, J. Yang, X. Luo, and P. Gong. Use of mobile phones in an emergency reporting system for infectious disease surveillance after the sichuan earthquake in china. *Bulletin of the World Health Organization*, 87(8):619–623, 2009.

[7] G. Yava, D. Katsaros, z. Ulusoy, and Y. Manolopoulos. A data mining approach for location prediction in mobile environments. *Data & Knowledge Engineering*, 54(2):121–146, 2005.