CDR-To-MoVis: Developing A *Mo*bility *Vis*ualization System From CDR Data

Manoranjan Dash [#1], Kee Kiat Koo [#2], James Decraene [*3], Ghim-Eng Yap [#4], Shonali Priyadarsini Krishnaswamy [#5], Wei Wu [#6], Joao Bartolo Gomes [#7], Amy Shi-Nash [*8], Xiaoli Li [#9]

[#] *Institute for Infocomm Research, A*Star, 1 Fusionopolis Way, Singapore 138632*
[1,2,4,5,6,7,9] {dashm, koo-kk, geyap, spkrishna, wwu, bartologjp, xlli}@i2r.a-star.edu.sg

[*] *R&D Labs, Living Analytics, Group Digital Life*
Singapore Telecommunications Limited, Singapore 239732
[3,8] {jdecraene, amyshinash}@singtel.com

***Abstract*—CDR (Call Detail Records) data is more easily available than other network related data (such as GPS data) as most telecommunications service providers (TSPs) maintain such data. By analyzing it one can find mobility patterns of most of the population thus leading to efficient urban planning, disease and traffic control, etc. But its granularity is low as the latitude and longitude (lat-lon) of a cell tower is used as the current location of all mobile phones that are connected to the cell tower at that time. Granularity can range between 10s of metres to several kms depending on population density of a locality. This is one reason why, although there are many existing systems on visualizing mobility of people based on GPS data, there is hardly any existing system for CDR.**

We develop a *Mo*bility *Vis*ualization System (MoVis) for visualizing mobility of people from their CDR records. First of all, given the CDR records of a user, we determine her stay regions (places where she stays for a significant amount of time). Trajectories of phone events (and lat-lon of cell towers) between stay regions are extracted as her trips. Start and end times of a trip are estimated using linear extrapolation. Based on the start and end times, temporal patterns are extracted. Trips with sufficient number of intermediate points are mapped to transport network that consists of train lines, bus routes and expressways. We use Kernel density estimation to visualize the most common path for a given origin and destination. Based on this we create a round-the-clock visualization of mobility of people over the entire city separately for weekdays and weekends. At the end we show the validation results.

I. Introduction

CDR data is generated and maintained by all telecommunications service providers. It shows the time stamp of a phone event and the lat-lon of the cell tower it is connected to. It is used primarily for generating billing records [1]. Analysis of CDR data has the potency of revealing mobility of the entire population thus leading to better public-related services such as urban planning, traffic control and disease control [2]. Because of its low cost, CDR data is more easily available than other location-based data, e.g. GPS data. But it lacks granularity because the location of a mobile phone is the lat-lon of cell tower it is connected to. Depending on the population density of a locality, granularity can vary from tens of metres to a few kms. With these contrasting characteristics of high *reachability* but low granularity, analysis of CDR data is interesting yet challenging.

In this paper we are concerned with extracting mobility information about a user from her CDR records. An important part of mobility is map-matching, i.e., mapping trajectories of phone events to transport network. There are many existing papers that show how to map GPS trajectories to transport network [3]. But GPS data is much more frequent and extremely regular compared to CDR records. High sampling rate and regular observations are key factors in achieving accurate map-matching. If a user does not make any phone event, CDR data may not show any record. Even though the user has made a trip, it does not show up in her CDR records. So, trips with few intermediate phone events may be rejected upfront.

Doyle et al [4] determines travel mode (train, bus) and major routes taken by mobile phone users. Their study involved CDR trajectories between Dublin and Cork which are more than 250km apart. The number of possible transport routes is very few. Another similar study addresses the problem of extracting intercity trips from CDR data and mapping the trips to intercity highways [5]. Our study involves users' mobility in a densely populated city like Singapore where travel distance generally do not exceed 15km. When travel distance is small, number of possible transport routes increases dramatically, and time to travel reduces significantly, thus reducing the number of intermediate phone events on the way.

There are several works that consider temporal and spatial population distribution using CDR data [2], [6]. These works do not distinguish between towers used by a user as her stay regions where she stays significant amount of time, and in-transit or intermediate points which are used when she is traveling. Furthermore, in these works the trajectories of phone events are not mapped to transport network. As we will soon see, mapping a trajectory to transport network requires specific departure time from origin and arrival time at the destination which are usully absent in the raw CDR data.

We develop a comprehensive *Mo*bility *Vis*ualization System (MoVis). Travel within a city, although of small distance and usually with few intermediate phone events, is repeated again and again. For example, a user goes from her home to office every weekday, or she goes to her favorite supermarket every

week, and so on. MoVis determines mobility patterns for individual users and for all users. It starts with determining stay regions of a user where s/he stays significant amount of time. Trajectories of phone events (and lat-lon of cell towers) between stay regions are extracted as her trips. Start and end times of a trip are estimated based on linear extrapolation. Trips with sufficient number of intermediate points are mapped to transport network that consists of train lines, bus routes and expressways. We use Kernel density estimation to visualize the most common path for a given origin and destination. Based on this we create a round-the-clock visualization of mobility of people over the entire city separately for weekdays and weekends. Finally, we validate the results using both direct and indirect approaches.

In the demo we will show these features of MoVis using a sample mobile network data of three months for 3.9 million users with (approx) 8 billion CDR records. Mobile network data is the service log when a mobile phone is connected to mobile network. It contains anonymised ID [1].

Unique contributions of this demo includes: (a) a method to extract stay regions from CDR data, (b) a method to extract trips a user makes within a city, (c) map-matching trajectories of CDR data to transport network, and (d) a round-the-clock visualization of trips of all users in a city.

The rest of the paper is organized as follows. Section 2 briefly describes the engine of MoVis. Section 3 discusses MoVis architecture. A walk-through of MoVis with screen shots of UI is presented in Section 4.

II. Methodology

See http://www1.i2r.a-star.edu.sg/∼dashm/supp.pdf for detailed descriptions of methods.

Determine Stay Regions and Trips: Consider a single user and her CDR records. A **stay region** is a location consisting of one or more cell towers that satisfy two thresholds, a temporal (Th) and another spatial (Sh). The user's records must show that she stayed in a stay region at least for Th time. A trajectory of phone events between any two stay regions is a **trip**. So, a trip consists of two stay regions as its start and end, and zero or more cell towers as intermediate points. Each trip is scanned once more to find out incidents of stay in a stay region with only one phone event or a sequence of phone events that does not satisfy Th. If time gap between the event(s) and its preceding or succeeding event is more than Th time, the trip is terminated in that stay region and a new trip is started that consists of the remaining phone events of the old trip. Note that this phase does not find any new stay regions. This phase is necessary because CDR data is not as frequent and regular as GPS data.

Home and Work Place Prediction: Among all stay regions home and work place are the most important as a user spends most of her time there. See [7] for further details.

[1]The anonymised ID is a machine generated ID via a two-step non-reversible AES encryption and hash process. There is no personal information about mobile subscribers in the data set, nor any content of calls or SMSs. The latitude and longitude in the dataset is at the mobile cell tower level, covering a range of 10s to 100s of meters.

Temporal Patterns: In order to find temporal patterns, we estimate start and end times for each trip. Consider a trip with at least two intermediate points, say a and b. Time(b) - Time(a) is required to cover Dist(a, b). Using linear extrapolation, we calculate *projected start time* and *projected end time*. If original start time is later than projected start time, retain the original. Similarly, if original end time is earlier than projected end time, retain the original.

Trip Distance: If we take the direct distance between the start and the end as the trip distance, we may underestimate it. On the other hand, if we calculate the trip distance by adding all consecutive cell tower distances (those cell towers appearing in the trip), we usually overestimate, particularly when there are many intermediate points. We use a map simplification algorithm (Sivalingam-Whyatt algorithm [8]). It can find out major changes in the shape of the trip curve, and keep only those intermediate points that retain the shape.

Representative Trip: There may be many trips between two stay regions of a user. Visualization of all such trips simultaneously appears messy. We applied a Markov Chain Model to show a representative trip. First of all, a graph like structure is created. If cell tower b appears in a trip immediately after a, then a and b are connected by an edge. The weight of the edge is the total number of times $a \rightarrow b$ appears in trips. Once the graph structure is ready, begin from $start$ node. Find the next node with the highest count having an edge from $start$. Repeat this procedure for the next node until end appears. The path from $start$ to end is the representative trip.

Map-Matching: Each trip is mapped to its nearest route. We included 5 train lines, 11 expressways, and 374 bus routes. Geometries for these routes are obtained from OpenStreetMap (openstreetmap.org). The entire Singapore map is divided into many Voronoi cells using cell tower lat-lon as seeds. R-Tree is used to find out the Voronoi cells intersected by a transport route. These Voronoi cells are stored in $RouteVoronoiCells$. A user's trip must have at least four intermediate points apart from $start$ and end. Voronoi cells intersected by the trip trajectory curve are stored in $TripVoronoiCells$. A transport route is selected in $ShortListedRoutes$ that has at least one common Voronoi cell between $RouteVoronoiCells$ and $TripVoronoiCells$. Distance is calculated between the user's trip to each route in $ShortListedRoutes$ by adding the perpendicular distance from each seed point in $TripVoronoiCells$ to the route. The route with the shortest distance is selected.

Most Common Route between an Origin and a Destination: Given an origin and a destination, first of all we collate all trips for all users. Frequencies of usages of cell towers are computed and fed to a two-dimensional Kernel density estimation (KDE) program. As a result, the most common path is prominently visible with high intensity color. The calculated Kernel densities of cell towers are fed to a global thresholding algorithm. This algorithm can automatically determine the threshold that separates the towers that appear in the prominently visible path from other cell towers.

Next we determine the sequence of these selected cell towers from origin to destination. This sequence is then mapped to its nearest transport route.

Tracking the Mobility of Population: Using KDE we can track the mobility of the population round-the-clock. There are two visualizations: (1) shows everything (both mobile and static), and (2) shows only the trips (not the static).

III. System Architecture

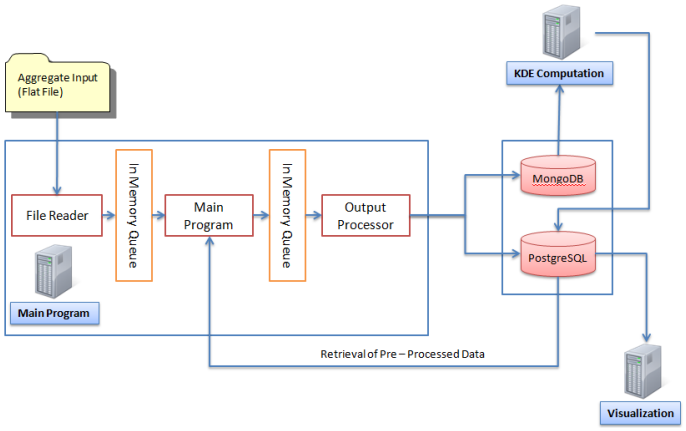

Fig. 1. MoVis Architecture: Multi-Thread Model

As illustrated in Figure 1, MoVis includes a Main Program (MP) that processes each user record one at a time. It extracts stay regions of the user, followed by her trips between pairs of stay regions. It computes projected start and end times of a trip. Based on this it finds temporal patterns. It can also show a representative trip using Markov Chain model. It maps individual trips to its closest transport route as part of map-matching. A Kernel Density Estimation (KDE) Program processes trips across the entire data set. It finds the most common route between a given origin and destination, and maps it to the nearest transport route. It also determines the round-the-clock mobility of the entire population. The Visualization Program shows the results of MP and KDE superimposed over the map of Singapore with the transport network in the background. We employ a variant of the Command Query Responsibility Segregation pattern in our design mainly for performance reasons (MongoDB as a write store and PostgreSQL / PostGIS as a read store) (http://msdn.microsoft.com/en-us/library/dn568103.aspx). A Map-Reduce version of the system is also implemented. There are also plans to move the data storage layer to Apache Hive.

IV. Demonstration

A user can interact with MoVis at individual level to visualize her stay regions, trips, temporal patterns of trips, representative trip, mapping of each trip to transport network, etc. She can also visualize commonly used routes between an origin and a destination, and mobility heat map of all people.

Stay Regions: Figure 2(left) shows top-5 stay regions as circles with Sh as radius. Percentage of time spent in a stay

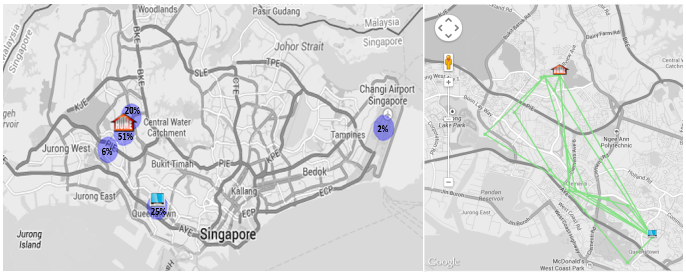

Fig. 2. First: Top-5 Stay Regions of a User; Second: Trips between Home-to-Work

region is shown inside the circle. She spends 36% of her time at home, and 16% at work, and so on. Figure 2(right) shows trips between home and work.

Representative Trip: In Figure 3 the left side shows many

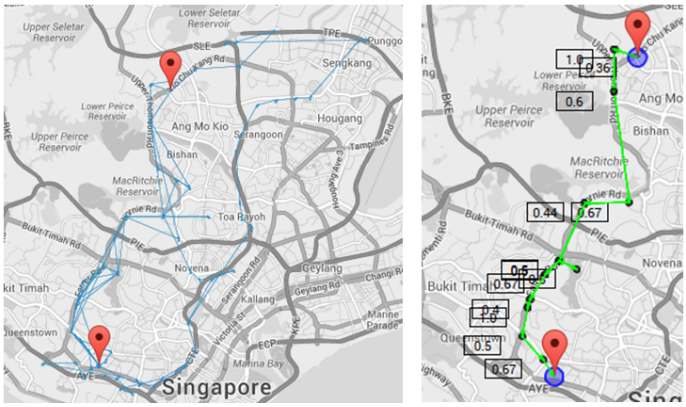

Fig. 3. Representative Trip Using Markov Model. Left: All trips, Right: Representative Trip

trips between two stay regions (in this case home and work palce). Many trips follows the same route while a few trips take different routes. A representative trip captures the most commonly used route. It (right side) shows the probabilities assigned to various intermediate points.

Temporal Patterns: Figure 4 shows the results of grouping start and end times for trips between home to work. Times are grouped into hourly slots starting with the earliest time. In this case the earliest start time of the user is 6:10AM.

Mobility Score: Number of stay regions, number of trips and distance per trip contribute to the mobility score of an individual. Mobility score of a planning area is the average mobility score of users staying there. Figure 5 shows that remote planning areas constitute the top-10 mobility scores whereas central business district (CBD) and nearby planning areas constitute the bottom-10.

Map-Matching If we include bus routes, most trips are mapped to a bus route because the number of bus routes far outnumber train lines and expressways. So, the system considers train lines and expressways separately from bus routes. Figure 6 shows the map of Singapore with train lines (blue) and expressways (grey). The actual trip (green) is

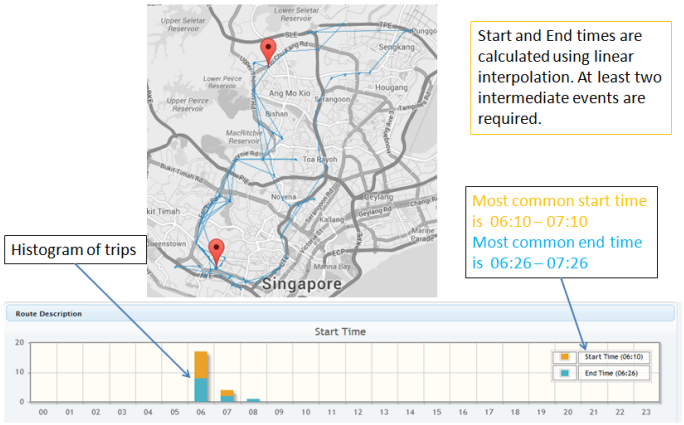

Fig. 4. Temporal Patterns: Start and End times of trips between home to work

Top Ten Planning Areas

Planning Area	Average Mobility Score
NORTH-EASTERN ISLANDS	0.71
CHANGI BAY	0.65
LIM CHU KANG	0.64
SIMPANG	0.60
SUNGEI KADUT	0.59
PUNGGOL	0.58
MARINA SOUTH	0.57
WOODLANDS	0.57
SEMBAWANG	0.57
CHOA CHU KANG	0.57

Bottom Ten Planning Areas

Planning Area	Average Mobility Score
ORCHARD	0.35
ROCHOR	0.36
DOWNTOWN CORE	0.38
NEWTON	0.39
MUSEUM	0.40
OUTRAM	0.41
NOVENA	0.42
KALLANG	0.43
RIVER VALLEY	0.43
TOA PAYOH	0.43

Fig. 5. Mobility Score by Planning Area

mapped to its nearest route, i.e., North-South train line. The corresponding train line segment with the start and end station names are highlighted.

Mobility Heat Map: In Figure 7(a-c) we show three screenshots at time (5, 8:30, and 10AM) of a round-the-clock mobility heatmap on weekdays. At 5AM the mobility is at the lowest. At 8:30AM, the train lines and expressways have become more prominent as people start going to work place. At 10AM morning peak hour is over, i.e., the mobility has reduced significantly. In other works authors have shown heatmap using all CDR records, but such heatmap misses significant change in traffic from 8:30 to 10AM which is captured here by considering only the trips. A user can set weekdays/weekends, bandwidth (for KDE), time window interval (1hr, 2hr), and increment (15 min, 30 min). Figure 7(d) shows heatmap for trips between an origin and destination.

Validation: Validation can be (a) direct (e.g. a sample survey) or (b) indirect (e.g. comparison with existing statistics). (a) In a survey of 20 participants, home prediction accuarcy is 100% (within 1 km), work place prediction is 94%, top-5 most frequent places is 92% and map-matching of top-5 most frequent trips is 86%. (b) Distance per trip is 8.4km compared to LTA (www.lta.gov.sg) statistics of 9.7km; number of trips per day per user is 2.29 compared to LTA's 2.35. Validation results show that MoVis is effective.

A video clip is in http://www1.i2r.a-star.edu.sg/∼dashm/ICDE15Demo.mp4.

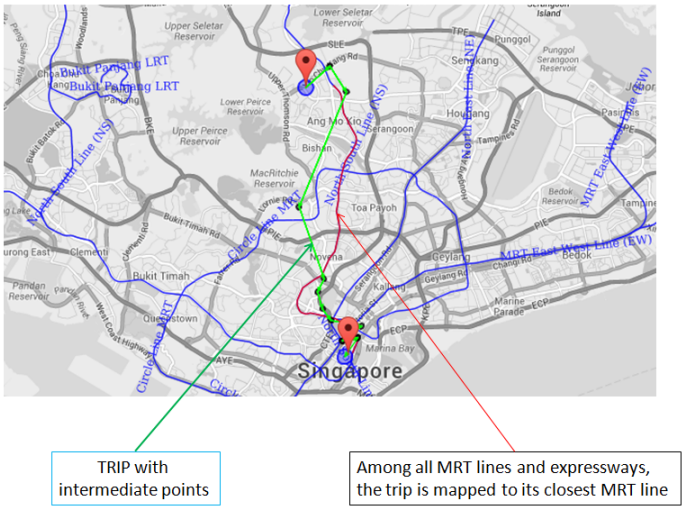

Fig. 6. Map-matching: Mapping a trip to transport network

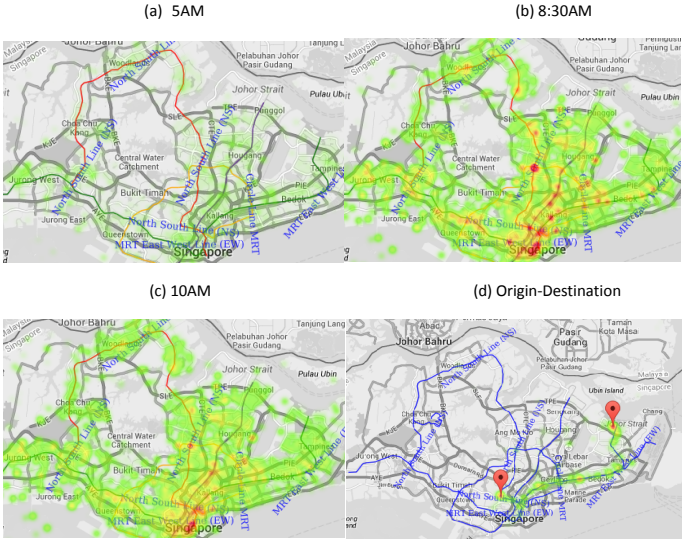

Fig. 7. (a)-(c): Round-The-Clock Heatmap on Weekdays of Mobility in Singapore; (d) Heatmap for trips between a given origin and destination

REFERENCES

[1] "Understanding call detail records," https://supportforums.cisco.com/document/53056/understanding-cdr-call-detail-records, Nov. 2010.

[2] A. Wesolowski and et al, "Quantifying the impact of human mobility on malaria," *Science*, vol. 338, no. 6104, pp. 267–270, 2012.

[3] S. Brakatsoulos and et al, "On map-matching vehicle tracking data," in *VLDB*, 2005, pp. 853–864.

[4] J. Doyle and et al, "Utilising mobile phone billing records for travel mode discovery," in *ISSC*, 2011, pp. 37–42.

[5] W. Wei and et al, "Studying intercity travels and traffic using cellular network data," in *NetMob*, 2013.

[6] C. Song and et al, "A data mining approach for location prediction in mobile environments," *DKE*, vol. 54, no. 2, pp. 121–146, 2005.

[7] M. Dash and et al., "An interactive analytics tool for understanding location semantics and mobility of users using mobile network data," in *MDM (Workshop)*, 2014, pp. 345–348.

[8] M. Visvalingam and J. Whyatt, "Line generalization by repeated elimination of points," *Cartographic Journal*, vol. 30, no. 1, pp. 46–51, 1993.