CDRoute: A Randomized Algorithm to Estimate Street-Level Commuting Routes Using Call Detail Records

Haewoon Kwak
Telefonica Research, Spain
kwak@tid.es

Edgar Olivares Mañas*
Universitat Politècnica de Cataluña, Spain
edgar@tid.es

Enrique Frias-Martinez
Telefonica Research, Spain
efm@tid.es

ABSTRACT

In this work we propose a randomized algorithm to estimate the street-level commuting route using call detail records (CDRs). Our algorithm works in three steps. First, for each individual, we create perturbations of estimated home and work locations within the cell tower coverage. Second, we identify a route between each pair of both locations through a web mapping service. Finally, we choose one route that maximizes the normalized number of commuting hour calls within certain distance from the route. We evaluate our algorithm with large-scale CDRs collected in two big cities during three months. The result shows that our algorithm finds, on average, 45.28%p and 35.58%p better routes than a naive approach without the perturbation step.

1. INTRODUCTION

Understanding human mobility has received much attention in recent years. The wide variety of projects trying to characterize human mobility cover such areas as studying its predictability [9], extracting frequent patterns [11], or predicting the next place based on movement history [1, 6].

For studying human mobility researchers have extracted data from various sources, such as GPS, traffic sensors, or location-based services. Among them, call detail records (CDRs) have become one of the major data sources because of three reasons: first, cell phones are pervasive in society; second, they capture large-scale human mobility data, especially when compared to GPS traces; and third, they have less biases toward popular spots than location-based services offering check-in features for venues, such as Foursquare. These CDRs advantages facilitate the estimation of home and work locations [5], places of interest [6] and the sequence of frequent visited locations. In CDRs, however, the location of an individual is usually expressed by the cell-phone tower that carried the communication. As a consequence, CDRs have two inherent limitations. One is that the location is captured only when a call takes place (i.e., low resolution); and the other is that the captured location (the tower used) is an approximation of the actual position (i.e., coarse granularity). These limitations make it hard to estimate the *street-level* commuting route between two frequent visited locations, such as home and work. The accurate estimation of street-level commuting routes, instead of the frequent sequence of cell towers or the probability distribution of placement, can bring us more sophisticated applications for location-based service, user profiling, target advertisement, traffic engineering, or urban planning.

In this work we propose a new randomized algorithm, called CDRoute, designed to estimate the street-level commuting routes of each individual between home and work locations using CDRs. Our algorithm works following three steps. First, for each individual, we create perturbations of estimated home and work locations within the cell tower coverage. Next, we identify a route between each pair of both locations through a web mapping service. Finally, we choose one route that maximizes the normalized number of commuting hour calls within certain distance from the route. We evaluate our algorithm with large-scale CDRs collected in two metropolitan areas during three months. The result shows that our algorithm finds, on average, 45.28%p and 35.58%p better routes than a naive approach without the perturbation step.

We note that our algorithm is not limited to CDRs but can work with any kind of geolocation datasets including GPS, hand-off patterns, or even user activities in location-based social networks.

2. RELATED WORK

In recent years, regularities in human mobility have been observed in various domains. The work by Song *et al.*, highlighted that a single person's location is predictable with 93% accuracy on average using CDRs [9]. Wang *et al.* found strong weekly and daily periodic movements over a few months [10], and even within daily life mobility, Mclnerney *et al.* was able to distinguish predictable states from unpredictable ones by using instantaneous entropy [7]. Related to that concept, Cho *et al.* discovered high spatial and temporal periodicity in short-ranged travel, while long-ranged travel was more explained by social relations [3]. This regularity gives a theoretical basis for the prediction of periodic movements.

Focussing on route and movement estimation, probabilistic approaches have been widely used in the literature. Görnerup proposed a scalable probabilistic method based on locality-sensitive hashing and graph clustering for inferring common routes from sequences of cells [4]. Saravanan *et al.* proposed to aggregate CDRs to find people's daily routes by constructing a Gaussian model that explained the probability of people being around specific towers [8]. Also, Isaacman *et al.* presented how to model people's movement in metropolitan-scale areas using spatial and temporal probability distribution observed from call detail records [6]. However, most of these approaches do not consider the geographic features when mining mobility.

*Work done while working at Telefonica Research

UrbComp '13 Chicago, USA

There are few studies to estimate the route people take at a street resolution. Becker *et al.* used cellular hand-off patterns to identify commuting routes [2]. They created a collection of hand-off patterns along each route by in-advance test driving. Their method achieves high accuracy but the dictionary of hand-off patterns for every route is essential in advance, being hard to be scaled for many cities and countries. By contrast, our algorithm does not require the additional data collection but uses globally deployed web mapping services for obtaining geographic information.

3. DATASET

Cell phone networks are built using a set of distributed cell towers, called base transceiver stations (BTS), that connect cell phones to the network. Each BTS gives cellular coverage to an area called a cell. A call detail record (CDR) is generated only when a mobile phone makes or receives a call (or an SMS or an MMS). For invoice purposes, the time when the call is made and the BTS tower that the mobile phone connects to are logged in a CDR. The tower location in CDR is then an approximation of the geographical position of a mobile phone at a given moment.

In this work we analyze fully-anonymized CDRs collected in 2009 during three months in two metropolitan areas, Madrid and Barcelona. No personal information was available for this study, and none of authors of this paper participated in the anonymization or extraction of the dataset. We focus our study in metropolitan areas because commuting routes typically take place within a urban area and cell towers in such areas are more densely located and cover smaller areas than in rural area. As a result of this finer resolution, we expect to characterize human commuting patterns more accurately. From all the data contained in CDRs, we use the anonymized identifiers of a caller and a callee, the time and date of the call, and the BTS towers used by a caller and a callee. From the whole dataset, the number of calls made during commuting hours of working days are 5 millions for Barcelona and 23 millions for Madrid.

4. RESEARCH GOAL

Our research goal is to estimate the street-level commuting route of an individual using CDRs. We reasonably assume that the daily commuting route is quite stable in the long term. We consider three requirements for this work. First of all, in order to estimate the commuting route, we should know where home and work locations are. Although many techniques are proposed to address this problem, we use the method introduced in [5] due to simplicity but high accuracy with large-scale CDRs. This method basically finds clusters of cell towers that are involved in the largest number of calls during home and work hours. We follow their method to detect home and work locations as it has been repetitively verified in different situations. Their method assigns a cluster of BTS towers as home and work locations, and the actual home and work can be located anywhere within the coverage area of each cluster of towers.

Second, we should consider geographical features between home and work locations. People do not follow the geodesic line between two locations but move along roads and streets. All these geographical features should be considered so as to estimate the street-level route between home and work. We will access this information using web mapping services.

Third, we should consider where calls are made during commuting hours. Intuitively, the route that is located near the location of more calls is more likely to be the commuting route than a route further way for those locations. For our particular study we will focus on the location of calls made or received during commuting hours. We define commuting hours as 7am to 10am and 5pm to 8pm on weekdays. We note that we omit Friday and also holidays for obtaining a more stable commuting route. Our commuting route estimation has to satisfy not only geographical features but also the electronic footprints people left in the form of CDRs.

As a result of above three requirements, the commuting route we discover connects two locations that are close to the home and work locations, satisfies geographical features, and maximizes the number of commuting hour calls within the certain distance from the route.

5. CDROUTE: DETECTING COMMUTING ROUTES WITH CDR

In this section we propose CDRoute, a street-level commuting route estimation algorithm using CDRs. It requires as inputs, for each individual, the home and work locations, and the locations of cell towers that handled calls during commuting hours. The algorithm consists of three steps: (1) creating perturbations of home and work locations within the tower coverage; (2) identifying a route between each pair of both locations using a web mapping service; and (3) choosing one that maximizes the normalized number of commuting hour calls within certain distance from the route. The algorithm outputs the most probable street-level commuting route with its associated means of transportation, such as walking, driving, or public transport.

5.1 Web Mapping Service

In CDRoute, a web mapping service is used for obtaining all geographical features between the home and work locations or its perturbed versions. For instance, Bing Maps or Google Maps, which offer a fine-grained navigation with a street-level resolution, are popular web mapping services. From the web mapping service we can obtain not only the detail route between two locations but also the duration and the distance of the route. Moreover, the web mapping service can find different routes according to the corresponding means of transportation.

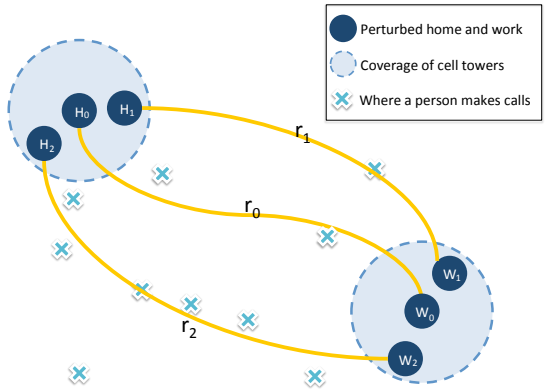
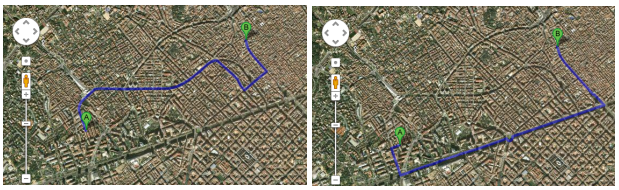


Figure 1: Overview of our proposed algorithm

5.2 Perturbations

For each individual, both home and work locations identified by [5] are expressed as longitude and latitude that represents the home and work clusters of cell towers. As we mentioned earlier, the real

home and work locations can be anywhere within the coverage area of the clusters of towers. To take these characteristics into account, we intentionally add noise, i.e. perturbations, to the home and work tower location. In the rest of the paper, we use *original* home to represent the home tower location. Within a circular area whose center is the latitude and longitude of a cell tower and a radius represents the coverage of the tower, we create N pairs of perturbed locations from the original home and work. In this work we conservatively assume the radius of the coverage as 200m in urban area. Figure 1 illustrates this process. H_0 and W_0 are the coordinates identified as home and work locations by [5], respectively, and H_1, W_1, H_2 and W_2 are perturbed home and work locations within the coverage of the cell tower. For each pair of perturbed home and work, we query the web mapping service and obtain the routes represented by the yellow lines.



(a) Original (b) Perturbed

Figure 2: Impacts of perturbations on routes

Figure 2 shows how small perturbations can bring a great impact on the route. Figure 2(a) presents the non-perturbed locations of two markers A and B. In Figure 2(b), a small perturbation of moving 100m to the south marker A has been applied, and we can observe the considerable variation in the route between marker A and B. By considering different means of transportation, such as public transport or walking, routes can be even more diverse.

5.3 Utility Maximization

Among the N routes provided by the web mapping service for the N pairs of perturbed home and work, we choose the most probable route by assessing the explanatory power of each route for the electronic footprints left in CDRs. For each route, r_i, we define the utility function, U_i, as the number of calls made during commuting hours within d meters from the route, c_i:

$$U(r_i) = \frac{c_i}{C} \tag{1}$$

where C is the total number of calls that the individual is involved in during commuting hours. In order to avoid the effect of bursty call behavior, we considered just one call per BTS tower per day. Whilst we can enhance the resolution of our route estimation by decreasing d, the straightforward choice is setting d as the diameter of the average coverage of a cell tower in an urban environment. After we finish the computation of the utility function for N routes, we then choose the most probable route, r_*, that maximizes the utility function U:

$$r_* = \underset{r}{\operatorname{argmax}} |U(r)| \tag{2}$$

To help understanding the algorithm, we computed U for the three routes in Figure 1: $U(r_0) = 2/10 = 0.2$, $U(r_1) = 1/10 = 0.1$, and $U(r_2) = 6/10 = 0.6$. We note that the sum of the utility functions does not necessarily adds up to 1. CDRoute will select r_2 as r_* because $U(r_2)$ is the maximum value.

6. RESULTS ANALYSIS

In this section we demonstrate how our algorithm improves the explanatory power of the estimated route from the perspective of the utility function U. As we mentioned, we experiment with CDRs collected from two major cities in Spain, Barcelona and Madrid, during three months and use two superscripts, ^{B} and ^{M}, to differentiate them. The experiments are conducted only for the individuals who made at least 10 calls during commuting hours for the period of time considered. For each individual we find home and work locations using [5] and generate 20 perturbations. We tried with different number of perturbations in both cities and found out that 20 is enough to capture the variability of routes and as a result the improvement of our algorithm. This number is also a refelction of the complexity of the geography in the areas under study.

Each perturbation is queried using the web mapping service with the three options of transportation: walking, driving, and public transport. We set d=200m, the average diameter of the coverage of a tower, as the proximity criteria that determines whether the call is near the route or not.

We begin with the comparison of the utility function U between r_0, the route estimated with the original, non-perturbed home and work, and r_*, the best route that maximizes U with a pair of perturbed home and work. We find out that the utility function for the best route for Barcelona, $U(r_*^B)$, (median: 0.7742, mean: 0.7433) is higher than that for the route for original home and work, $U(r_0^B)$ (median: 0.0, mean: 0.2907). Surprisingly, the median of $U(r_0^B)$ is zero; which indicates that the route between non-perturbed home and work, in general, is not even close to where calls are made during commuting hours. This finding reveals the prominent role of the perturbations for the route estimation. The difference of medians is significant and confirmed by a Mann-Whitney's U test (U=676,081, $p < 0.001$). Similarly, for Madrid we observe higher U for the best route (median: 0.6667, mean: 0.6474) than the original one (median: 0.250, mean: 0.2918), and it is statistically significant (U=419,780, $p < 0.001$).

In order to assess the improvement on the route estimation, we define the improvement I as the difference between $U(r_*)$ and $U(r_0)$. Since U is the normalized index from 0 to 1, I directly measures what percentage of calls *additionally* support the route. We compute I for each case and depict the cumulative distribution for both cities in Figure 3. The median of I^B and I^M are 0.4286 and 0.2759, respectively. In other words, our algorithm estimates the commuting route with at least 42.86%p (percent point) better for 50% of cases for Barcelona and 27.59%p for Madrid with regard to U. Furthermore, except for 7.29% and 7.11% for Barcelona and Madrid, respectively, our algorithm can always find a better route that is supported by more calls than the original non-perturbed one. This great improvement also implies the home and work locations identified by [5] are quite acceptable. We note that the average of I^B and I^M is 0.4528 and 0.3558, respectively.

Finally, we compare the length of r_0 with r_*. This result quantitatively shows how the route can change by adding small perturbations and by considering different means of transportation. We highlight that our route is not intentionally over-fitted to where calls are made but optimally found from the set of recommended routes obtained from perturbations, and as a result, the selected route can be longer than the original one. We observe that r_* is slightly longer than r_0 for the two cities (median Δ^B: 0.40km, mean Δ^B: 0.77km, median Δ^M: 0.30km, mean Δ^M: 1.01km). Surprisingly, however, 34.57% of r_*^B and 14.64% of r_*^M are shorter than or equal

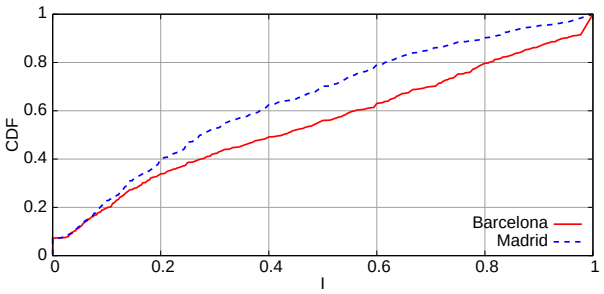


Figure 3: Cumulative distribution of I

to r_0^B and r_0^M, respectively. Moreover, 89.93% and 81.90% of them have higher U than the corresponding r_0^B and r_0^M, respectively. In this case, even though the route is shorter, it is supported by more calls. This becomes an indication that the benefit of our algorithm does not come from the lengthened distance but the appropriate placement of the route.

If the means of transportation, such as walking, driving, and public transport, are considered, we observe a significant impact on the distance of r_* from Barcelona confirmed by a Kruskal Wallis test ($\chi^2(2)^B = 90.3968$, p^B <0.001), but we cannot find the statistically significant impact from Madrid ($\chi^2(2)^M = 3.3709$, p^M >0.05). For Barcelona, the post-hoc test using Mann-Whitney tests with Bonferroni correction shows the significant difference of the distance of r^* between driving and public transport (p <0.001, r(the effect size)=0.229), driving and walking (p <0.001, r=0.10), and between walking and public transport (p <0.01, r=0.09). In order to understand the difference between two cities we require the knowledge of where residential areas are and how urban transportation systems are designed for both cities. This leaves for future work.

7. DISCUSSIONS AND SUMMARY

In this work we have presented a randomized algorithm, CDRoute, that estimates the most probable commuting route at a street-level using CDRs. The algorithm assumes that users have a stable commuting route over time and a minimum level of calls during commuting hours. Our experiments with CDRs collected in two big cities during three months demonstrate how our proposed algorithm improves the explanatory power of the estimated route thanks to the perturbation of home and work locations. Also, CDRoute can be potentially applied to any dataset that contains location information.

Verifying our estimated commuting path with the ground-truth data becomes a new research challenge. In that sense we can have two complementary approaches: synthesized call detail records [6] and survey data from a small group of people. Using the first approach we plan to create virtual individuals who have random home and work locations using the available residential and industrial information of an urban area. With this information, we can obtain their reasonable commuting routes by web mapping services with its associated mean of transportation. This information defines the ground truth data. By assuming that individuals make a call on their commuting route, we synthesize call detail records with some parameters reflecting temporal characteristics. Then we planb to apply CDRoute to these synthesized CDRs and validate the accuracy of the algorithm. Collecting survey data could be an alternative solution. Participants would be asked to make a note about their commuting route and where they use mobile devices in commuting hours. Moreover, with these approaches we can answer some questions about the performance and the robustness of the algorithm, such as how many call detail records are required for recognizing the commuting route accurately.

8. REFERENCES

[1] A. Asahara, K. Maruyama, A. Sato, and K. Seto. Pedestrian-movement prediction based on mixed markov-chain model. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 25–33, 2011.

[2] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. Route classification using cellular handoff patterns. In *Proceedings of the 13th international conference on Ubiquitous computing*, UbiComp '11, pages 123–132, 2011.

[3] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1082–1090, 2011.

[4] O. Görnerup. Scalable mining of common routes in mobile communication network traffic data. In *Proceedings of the 10th international conference on Pervasive Computing*, Pervasive'12, pages 99–106, 2012.

[5] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Identifying important places in people's lives from cellular network data. In *Proceedings of the 9th international conference on Pervasive computing*, Pervasive'11, pages 133–151, 2011.

[6] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, MobiSys '12, pages 239–252, 2012.

[7] J. Mclnerney, S. Stien, A. Rogers, and N. R. Jennings. Exploring periods of low predictability in daily life mobility. In *Proceedings of Workshop on Mobile Data Challenge by Nokia*, 2012.

[8] M. Saravanan, S. Pravinth, and P. Holla. Route detection and mobility based clustering. In *Proceedings of IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application (IMSAA)*, pages 1–7, 2011.

[9] C. Song, Z. Qu, N. Blumm, and A.-L. BarabÃąsi. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[10] Z. Wang, M. A. Nascimento, and M. MacGregor. On the analysis of periodic mobility behavior. In *Proceedings of Workshop on Mobile Data Challenge by Nokia*, 2012.

[11] R. Xie, Y. Ji, Y. Yue, and X. Zuo. Mining individual mobility patterns from mobile phone data. In *Proceedings of the international workshop on Trajectory data mining and analysis (in conjunction with UbiComp 2011)*, TDMA '11, pages 37–44, 2011.