Human Trajectory Recovery via Mobile Network Data

Guangshuo Chen, Aline Carneiro Viana, Marco Fiore

Human Trajectory Recovery via Mobile Network Data

Guangshuo Chen[1,2], Aline Carneiro Viana[2] and Marco Fiore[3]

[1]*École Polytechnique, Université Paris-Saclay, France*
[2]*INRIA, Université Paris-Saclay, France*
[3]*CNR - IEIIT, Italy*

For human mobility studies across many disciplines, mobile network data serves as a primary source of human footprints with geo-referenced and time-stamped records of human communication activities. Nevertheless, the quality of mobility information provided by mobile network data is usually not satisfactory on many users. Due to the nature of human communications, individual trajectories inferred from mobile network data are often substantially incomplete, and the pattern of missing locations is not uniform over time but is highly related to communication activities. In this paper, we propose a novel hierarchical approach based on tensor factorization to reconstruct such incomplete individual trajectories. The data-driven simulation shows that, with ground-truth locations in precision of 200 meters, our approach can recover a trajectory from 10% of its known locations with a distance error below 750 meters, which outperforms the existing proposals in the literature.

Keywords: Trajectory completion;movement inference;data enrichment;operator-collected datasets

1 Introduction

Human footprints are essential to human mobility investigation by providing information of movements. Datasets of CDR (namely, charging data records or call detail records) have been considered as a primary data source of human footprints [1]. Collected by cellular network operators as necessary information for billing purposes, CDR cover large populations easily and contain data entries about when, where and how mobile network subscribers have issued and received voice calls, sent and received text messages, or established data traffic sessions. Such mobility information of CDR is needed by many research communities [1]. Nevertheless, human communication activities are heterogeneous and usually sparse in time [2]. When forging time-stamped and geo-referenced CDR into trajectories with a stable temporal resolution, not all time slots of a trajectory have locations captured. Due to this missing location problem, many works cannot utilize CDR smoothly as readily trajectories, but have to first deal with lacking sufficient mobility information (*e.g.*, in [3]). In this context, we propose a novel hierarchal approach aiming at reconstructing trajectories using mobile network data. To the best of our knowledge, only one approach based on linear and cubic interpolation is proposed explicitly for this application scenario [4] in the literature[†]. We employ a large-scale mobile network dataset that provides locations with precision of 200 meters to evaluate the performance of our approach comparatively. The results show that our approach outperforms the existing one and can generate reconstructed trajectories with an average distance error below 750 meters from 10% of mobility information.

2 CDR completion problem

Our target is CDR-based trajectories with a stable temporal resolution of one hour. Given a user u, his trajectory is a time series of locations, represented as $L^u_{\mathcal{T}} = \{\mathbf{l}^u_i | i \in \mathcal{T}\}$, where $\mathbf{l}^u_i$ is the representative location (that the user spends the most time) in the i-th time slot, and $\mathcal{T}$ is a set of N equivalent time slots

[†] We refer the reader to [5] for a more elaborate version of the state of the art.

covering the observing period. We use an *observation set* as $\Omega_u \subseteq \mathcal{T}$ to represent the time slots in which the representative locations are observed, and thus have the observed part of the trajectory, represented as $L^u_{\Omega_u} = \{\mathbf{l}^u_i | i \in \Omega_u\}$. Besides, we employ 2-dimensional Euclidean space to represent locations: each location $\mathbf{l}^u_i$ is a 2-dimensional coordinate. Our *CDR completion problem* is to infer all missing locations in a trajectory $L^u_{\mathcal{T}}$ as precise as possible, *i.e.*, mathematically,

$$\min \sum_{i \in \Omega^C_u} |\mathbf{l}^u_i - \hat{\mathbf{l}}^u_i| / |\Omega^C_u|, \quad s.t.\ L^u_{\Omega_u} \tag{1}$$

where $\Omega^C_u = \mathcal{T} - \Omega_u$ and $\hat{\mathbf{l}}^u_i$ is the estimated location of the i-th time slot.

3 Context-enhanced CDR completion

To address the CDR completion problem, we propose a hierarchical approach that leverages the user's context of natural sleeping cycles and redundant mobility patterns. The approach receives an incomplete trajectory $L^u_{\Omega_u}$ as input and infers its missing locations $\hat{L}^u_{\Omega^c_u}$ as output. It is composed of two steps, introduced in the following.

3.1 Nighttime data enhancement

This first step is to address the trajectory's heterogeneity. Due to the nature of human communication activities, the captured locations are not uniform over time. As illustrated in Fig. 1(a), there are usually far more locations captured during daytime hours than nighttime. Therefore, we employ this step to fill only nighttime time gaps in the trajectory with the user's home location identified. We leverage human sleeping cycle and adopt our `stop-by-spothome` strategy proposed in [6][‡] to identify a user's nighttime $(10pm, 7am)$ locations. The strategy determines a nighttime period that a user is most likely to stay at home adaptively. For the strategy, we refer the reader to [6] for more detail. Accordingly, we fill the time slots in this period with the identified home location (*i.e.*, the most frequent location during nighttime), so as to lighten nighttime data loss.

3.2 Temporal improved data completion

Hereby we infer all the remaining missing locations on the basis of locations captured by CDR and inferred in the first step. As an example of a raw CDR-based trajectory given in Fig. 1(a), we see a common human mobility feature, *i.e.*, daily locations are highly repetitive in the trajectory. Thus, we organize the trajectory in a structural form and employ *tensor factorization*, *i.e.*, a specific technique to recover redundant structural data [8], for the location inference. The repetitive pattern of human movements exists on hourly, daily, and weekly basis. We divide the observing period into one-week sub-periods that contain one-day sub-trajectories, and then convert the trajectory $L^u_{\mathcal{T}}$ to a three dimensional tensor $\mathcal{X}^u$ as shown in Fig. 1(b). The tensor $\mathcal{X}^u = \left\{X^u_{ij}\right\}_{n_w \times n_d}$ is composed of each one-day sub-trajectory X^u_{ij} in the j-th day of the i-th week, where n_w is the number of weeks and n_d is the number of observing days in each week. Mathematically, the tensor $\mathcal{X}^u \in \mathbb{R}^{p \times q \times r}$, where $p = n_w$, $q = n_d$, and $r = 2N/(n_d n_w)$, has known values within the indices of the observation set Ω_u.

Now we construct the optimization problem for recovering missing values in the tensor $\mathcal{X}^u$. For such inference to be applicable, we assume that the tensor $\mathcal{X}^u \in \mathbb{R}^{p \times q \times r}$ has a CPD (canonical polyadic decomposition) [8] of three d-rank metrics $A \in \mathbb{R}^{p \times d}$, $B \in \mathbb{R}^{q \times d}$, and $C \in \mathbb{R}^{r \times d}$. In the CPD, each value $\mathcal{X}^u_{ijk}$ in the tensor is approximated as $\mathcal{X}^u_{ijk} = \sum_{\delta=1}^{d} A_{i\delta} B_{j\delta} C_{k\delta}$. For simplicity, we employ the concise CPD expression, used by Kolda *et al.* [8], *i.e.*, $\mathcal{X}^u = [\![A, B, C]\!]$. Leveraging the CPD, the incomplete tensor $\mathcal{X}^u$ can be recovered as $\hat{\mathcal{X}}^u = [\![\hat{A}, \hat{B}, \hat{C}]\!]$ by solving the following TF (tensor factorization) problem:

$$(\hat{A}, \hat{B}, \hat{C}) = \underset{A,B,C}{\arg\min} \sum_{(i,j,k) \in \Omega_u} (\mathcal{X}^u_{ijk} - [\![A, B, C]\!]_{ijk})^2 + \lambda(||A||^2_F + ||B||^2_F + ||C||^2_F), \tag{2}$$

[‡] Major content of [6] is also presented in [7].

where λ is a penalty parameter to avoid overfitting and $||\cdot||_F$ is the Frobenius norm. By solving this problem, we obtain an approximation $\hat{\mathcal{X}}^u$ and accordingly, the inference of missing locations in $L^u_{\mathcal{T}}$.

Nevertheless, Eq. (2) is a standard TF problem which treats each dimension of the tensor equally. Regarding human mobility redundancy, daily repetitive patterns are usually stronger than weekly ones; also, locations of consecutive time slots may be identical. To have a realistic and accurate location inference, we induct two more constrains into the optimization problem Eq. (2) using Toeplitz matrices as illustrated in Fig. 1(c). The first constrain is to emphasize the daily repetitive pattern. For that, we construct a matrix $D = \text{Toeplitz}(1,-1,0)_{q\times q}$ (*i.e.*, a matrix with central diagonal given by 1, and the first upper diagonal given by -1, and the others given by 0), and induct into Eq. (2) a constrain $||\mathcal{X}^u \times_q D||_F^2$, where $\times_q$ is the *tensor-matrix* product [8] (*i.e.*, $(\mathcal{X}^u \times_q D)_{imk} = \sum_{n=1}^{q} \mathcal{X}^u_{ink} D_{mn}$) on the second dimension of the tensor $\mathcal{X}^u$. Intuitively, $||\mathcal{X}^u \times_q D||_F^2$ represents the sum of squared differences of location coordinates in the same hour of consecutive days. Combining CPD, we have $(\llbracket A,B,C \rrbracket \times_q D)_{imk} = \sum_{n=1}^{q}\sum_{\delta=1}^{d} A_{i\delta}B_{n\delta}C_{k\delta}D_{mn} = \sum_{\delta=1}^{d} A_{i\delta}(\sum_{n=1}^{q} D_{mn}B_{n\delta})C_{k\delta} = \sum_{\delta=1}^{d} A_{i\delta}(DB)_{m\delta}C_{k\delta}$ and equivalently, $\llbracket A,B,C \rrbracket \times_q D = \llbracket A,DB,C \rrbracket$, and import into Eq. (2) is $||\llbracket A,DB,C \rrbracket||_F^2$ as the first constrain. Next, to enhance the similarity between consecutive time slots, similarly, we make another Toeplitz matrix $H = \text{Toeplitz}(1,0,-1,0)_{r\times r}$ and construct the second constrain as $||\mathcal{X}^u \times_r H||_F^2 = ||\llbracket A,B,HC \rrbracket||_F^2$. Accordingly, our novel tensor factorization problem with the two constrains is constructed as follows:

$$
\begin{aligned}
(\hat{A},\hat{B},\hat{C}) = \arg\min_{A,B,C} & \sum_{(i,j,k)\in\Omega_u} (\mathcal{X}^u_{ijk} - \llbracket A,B,C \rrbracket_{ijk})^2 \\
& + \lambda(||A||_F^2 + ||B||_F^2 + ||C||_F^2) \\
& + \lambda_D ||\llbracket A,DB,C \rrbracket||_F^2 + \lambda_H ||\llbracket A,B,HC \rrbracket||_F^2
\end{aligned}
\tag{3}
$$

where λ_H and λ_D are parameters avoiding overfitting. Practically, the problem in Eq. (3) is solved (instead of Eq. (2)) to obtain $\hat{\mathcal{X}}^u$. It is a combination of multiple least square problems, and thus, can be addressed by ALS (alternating least squares) [8].

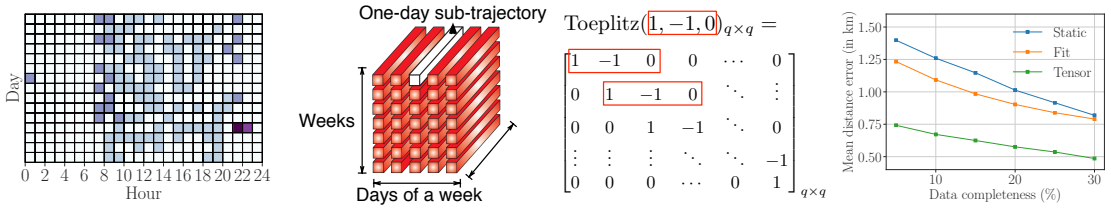
Overall, the tensor factorization infers all the remaining missing locations. The last operation in this step is to extract from the tensor approximation $\hat{\mathcal{X}}^u$ the complete version of the slotted CDR-based trajectory $L^u_{\mathcal{T}}$. We use $\hat{L}^u_{\Omega^C_u}$ to represent the set of those inferred locations for the missing time slots, *i.e.*, $\hat{L}^u_{\Omega^c_u} = \{\hat{\mathbf{l}}^u_i | i \in \Omega^c_u\}$.

4 Performance evaluation

We employ a data-driven simulation to evaluate our approach. The dataset that we leverage is collected from Shanghai, by a major cellular network operator in China. It contains trajectories of 28K mobile network subscribers generated during 10 weekdays of two consecutive weeks, with a temporal resolution of one location per hour. Each trajectory has at least 20 locations observed per day and has $10 \times 24 = 240$ one-hour time slots in total. In the simulation, we compare our approach with the other two proposals. The `static` approach is to fill all the empty time slots with the nearest preceding or ascending locations. The `fit` approach, proposed by Hoteit *et al.* [4], is to infer missing locations by linear interpolation on location coordinates if a user has a daily radius of gyration less than 3 km or cubic polynomial interpolation if a user has a larger radius of gyration. It is worth noting that the `fit` approach is the only one proposed explicitly for the CDR completion problem.

The procedure of the simulation is designed as follows. *(i)* Duplicate each trajectory in the ground-truth datasets to "mimic" slotted CDR-based trajectories observed in 60 days. *(ii)* Generate the target incomplete slotted CDR-trajectories with the completeness percentage of $\{5,10,15,20,25,30\}$. Note that we employ the per-hour inter-event distributions of real voice calls so as to let these target trajectories follow actual loss patterns in locations captured by CDR. We refer the reader to [6] for the information of the distributions. *(iii)* Apply the baseline `static`, `fit` techniques as well as our approach on the target CDR-based trajectories, and then obtain their inferred complete versions. *(iv)* For each slotted CDR-based trajectories inferred from the techniques, compute the performance metrics introduced later on.

We compute the distance error as the average distance between the estimated and actual locations on all time slots having unknown locations. Mathematically, given a CDR-based trajectory $L^u_{\mathcal{T}}$ with a loss set Ω^C_u,



(a) (b) (c) (d)

Figure 1: (a) Data loss in an actual CDR-based trajectory: each block represents the representative location of an one-hour time slot; each color represents an observed cell tower location while white color represents *missing*. (b) An illustration of converting a trajectory $L^u_{\mathcal{T}}$ into a tensor $\mathcal{X}^u$. (c) An example of the Toeplitz matrix. (d) Distance error on average versus completeness across trajectories of the `Shanghai` dataset.

the distance error of the trajectory is computed as follows: $\text{error}(\Omega^c_u, L^u_{\mathcal{T}}) = \frac{1}{|\Omega^c_u|}\sum_{i\in\Omega^c_u} ||\mathbf{l}^u_i - \hat{\mathbf{l}}^u_i||_{\text{geo}}$, where $\mathbf{l}^u_i$ and $\hat{\mathbf{l}}^u_i$ represents the actual and estimated locations at the i-th time slot respectively. Fig. 1(d) show the mean distance error of all completed trajectories in each ground-truth dataset, where our approach is marked by `tensor` and is compared with `static` and `fit`. We can clearly observe the following. The distance error of the approach is less than the ones of other comparison techniques. When the trajectory completeness $\geq 10\%$, the approach can almost have the distance error below 0.75 km. It is worth noting that the size of area that the location represents $200m \times 200m$ in the latter. Such distance error is relatively good regarding the location resolution of the ground-truth datasets. The distance error decreases with the increasing of data completeness, and moreover, the differences among the techniques become smaller with the same increasing. This indicates that the increasing of mobility information contributes to all the techniques. Still, the distance errors of the `fit` and `static` approaches are higher than the ones of the rest, indicating that utilizing the redundancy of human mobility helps to the completion a lot, particularly when the completeness is low. Overall, these results support the advantage of our approach over the other competitors.

References

[1] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 124–161, 2016.

[2] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *IEEE INFOCOM 2011*, pp. 882–890, IEEE, Apr. 2011.

[3] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, pp. 1018–1021, Feb. 2010.

[4] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Computer Networks*, vol. 64, pp. 296–307, May 2014.

[5] G. Chen, S. Hoteit, A. C. Viana, M. Fiore, and C. Sarraute, "Individual Trajectory Reconstruction from Mobile Network Data," Technical Report RT-0495, INRIA Saclay - Ile-de-France, Jan. 2018.

[6] S. Hoteit, G. Chen, A. C. Viana, and M. Fiore, "Filling the gaps: On the completion of sparse call detail records for mobility analysis," in *ACM CHANTS 2016*, pp. 45–50, ACM, Oct. 2016.

[7] S. Hoteit, G. Chen, A. C. Viana, and M. Fiore, "Spatio-Temporal Completion of Call Detail Records for Human Mobility Analysis," in *CoRes*, May 2017.

[8] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, pp. 455–500, Aug. 2009.