Takeaways in Large-scale Human Mobility Data Mining

(Invited Paper)

Guangshuo Chen*†, Aline Carneiro Viana† and Marco Fiore‡
*École Polytechnique, Université Paris Saclay, France, guangshuo.chen@inria.fr
†INRIA, Université Paris Saclay, France, aline.viana@inria.fr
‡CNR - IEIIT, Italy, marco.fiore@ieiit.cnr.it

***Abstract*—Employing mobile devices to perform data analytics is a typical fog computing application that utilizes the intelligence at the edge of networks. Such an application relies on the knowledge of the mobility of mobile devices and their users to deploy computation tasks efficiently at the edge. Therefore, this paper surveys the literature on the mobility-related utilization of operator-collected CDR (charging data records) – the most significant proxy of large-scale human mobility studies. We provide an innovative introductory guide to the CDR data preliminary and reveal original issues regarding CDR-based mobility feature computation and applications at the edge. Our survey plays important roles in investigating both human mobility and fog computing involving mobile devices.**

I. Introduction

The proliferation of mobile devices at the edge of cellular networks brings the possibility of collecting large-scale human behavioral data [1]. In the past decade, mobile devices have become the most popular data source for investigating human behavior or related issues [2], such as social relations [3], network traffic [4], and human mobility [5, 6].

Meanwhile, recent advances in mobile devices and mobile operating systems make it possible to employ mobile devices as data processing nodes rather than human behavior sensors. Applying distributed data analytics at the edge of cellular networks allows conducting data collection and processing more efficiently and securely [1], which alleviates heavy computation and storage pressure and also resolves data privacy concerns as in centralized data processing [7, 8].

To this end, it is essential to understand the behavior of mobile users in the network, particularly their mobility, to conduct intelligent utilization of network resources at the edge. The knowledge of such behavior helps to understand *where mobile devices are located* and consequently, *where and when their resources can be leveraged*. Therefore, it is necessary to study the way human behaves regarding habits of mobility, what will drive the spatiotemporal availability of mobile devices playing as both resource consumers and providers in fog computing.

Mobile devices, having their roles in fog computing, are also service consumers in cellular networks. Therefore, the mobility of mobile users can be obtained and investigated by leveraging operator-collected mobile phone records, or namely CDR (*charging data records)* [9]. Nowadays, collecting CDR is the most common mean of acquiring human behavioral data, which can easily cover broad areas and user groups with minimal cost [2]. Accordingly, CDR datasets are often employed in human mobility studies, bringing as main advantage large-scale populations and long observing periods [2].

This paper reviews the literature on the CDR data utilization for human mobility studies. Due to the fact that the quality of CDR-based mobility data is determined by the nature of human communications and varies widely across users, both data preliminary and processing need to be carefully designed and implemented. Yet the description of the data preliminary is sometimes neglected in research works, which questions the validity of their results and conclusions. Hence we summarize in this paper the common practices and our experience on dealing with mobility data extracted from CDR datasets and provide the significant takeaways in terms of data preliminary, mobility feature computation, applications, and future research directions.

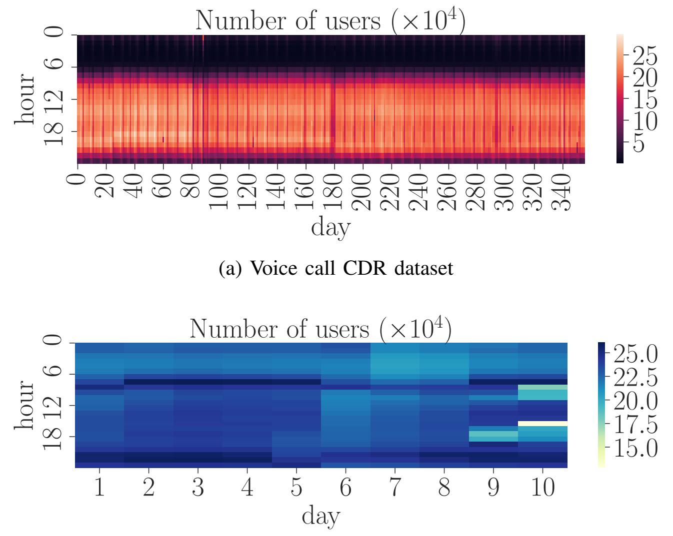
Our survey differs from the previous literature reviews of human mobility or network traffic analyses. They either summarize the models, applications, and techniques that are designed or employed for characterizing and utilizing human mobility as in [10, 11] or cover the vast literature of multiple research communities on mining mobile phone records as in [2]. Instead of the "outcomes" that are originated from CDR or other mobility data proxies, this paper mainly focuses on how they conduct the data preliminary and processing on CDR (via, *e.g.*, concepts, methodologies, and techniques) to obtain reliable and convincible results. We believe that our discussion in this paper, summarized as the takeaways regarding CDR data mining, will provide direct and valuable guidance to those who are working on mobility data.

II. Collecting human mobility data

A. Telecommunication events and their CDR

The availability of mobility data is the most fundamental requirement for human mobility analyses. In the literature, CDR, generated by mobile devices and collected by cellular operators, are the primary choice among a variety of mobility data proxies (CDR, WiFi, GPS, and travel surveys). CDR describe mobile devices' telecommunication events and are usually time-stamped and geo-referenced so that they can be leveraged as human mobility data [2]. Moreover, as the necessary data of cellular operators for billing or network management purposes, they are collected in a very large population at a small cost.

In the 3GPP lexicon, a series of CDR types are defined corresponding to telecommunication events such as voice calls, text messages, internet visits, mobility updates, and location requests [9]. Nevertheless, not all of them contribute to research works: only voice call and text message CDR are commonly seen [2], with a large and growing body of the literature on the their corresponding events (*e.g.*, [12]) and the mobility data extracted from them (*e.g.*, [13, 14]). It is known that such mobility data has limited spatiotemporal granularity and also suffers from a high degree of temporal heterogeneity and sparsity [15]. Internet visit CDR appear in a few of human mobility studies. They provide better mobility data with higher



(a) Voice call CDR dataset

(b) Internet visit CDR dataset

Fig. 1: Heatmaps of the number of users of our (a) voice call and (b) internet visit CDR dataset.

spatiotemporal granularity than voice call CDR and thus are often used as the ground-truth of the latter as in [13], while they are mostly collected within short observing periods [2].

B. Real-world CDR dataset

Hereby we give an overview of the two most important CDR datasets employed by our mobility analyses, as an example of real-world CDR datasets.

The first one consists of voice call CDR generated by 3.6M prepaid mobile subscribers in Mexico during approximately one-year. Each voice call CDR contains the caller's and callee's hashed identifiers, the call duration, the call initial timestamp, and the location of the cell tower to which the caller's device is connected to when the call originates. We portray in Fig. 1a the heatmap of the number of users in each hour of the voice call CDR dataset. We see clearly daily and weekly repetitive patterns of voice call behavior: there is an active period of making phone calls on each day and these active periods of each week are quite similar.

The second one consists of internet visit CDR of 0.6M of mobile devices in Shanghai while we only have the timestamps and cell tower locations because the other critical CDR attributes are removed by the data provider for privacy reasons. Similarly, the number of users per hour is illustrated by the heatmap in Fig. 1b. We see that there is still a daily repetitive pattern of internet visits but is less heterogeneous than voice calls, meaning that these CDR can provide more complete mobility data with less temporal sparsity.

It is worth noting that we observe imperfectness on both heatmaps: the user numbers of certain hours are significantly less than the others. This can be explained by public holidays and data collection abnormalities. Such imperfectness is common in CDR datasets and brings the necessity to perform data preliminary, *i.e.*, selecting appropriate populations of study, observing periods, and CDR attributes. To this end, we integrate our experience on data preliminary and summarize the common practices used in the literature, introduced next.

III. Common practices in mobility data processing

The reliability of data preliminary determines the quality of data mining and the representativeness of results obtained. In our experience of mining our CDR datasets, we apply a multitude of data preliminary steps and study the start-of-the-art works having detailed data preliminary description. In this section, we summarize common and effective practices in terms of data preliminary for CDR-based human mobility investigation.

- **Extract location coordinates via third-party services**. Locations are usually inherent as GPS coordinates in CDR while sometimes they appears as their original form, *i.e.*, Cell ID, and their coordinates need to be extracted manually. In this case, certain third-party services are available, including OpenCelliD[1], France OpenData[2], Google Geolocation[3], Unwired Labs[4], OpenSignal[5] and Mozilla Location Service[6]. They are usually powered by community databases and should be chosen carefully according to both their areas of study and data contributors.
- **Filter out "bad" users**. It is common to select users of study for reliability or generality by setting corresponding thresholds. Although there is hardly a standard, a relatively common threshold for voice call CDR is to keep those who have ≥ 0.5 Call/Hr and unique locations $N_L \geq 2$ as in [5, 13, 16], which can keep significant user locations and sufficient mobility information [13]. Note that such filters may drop a large number of users and should only be applied on mobility data having a large population.
- **Reduce temporal/spatial resolutions**. A fairly good setting on resolutions can reduce data quality requirement. For temporal resolutions, depending on CDR types and data quality, 15 minutes [13], 1 hour [5, 16], and 2 hours [17] are common. For spatial resolutions, a common practice is to merge adjacent locations via clustering methods (*e.g.*, DBScan, Optics), as in [15, 17].
- **Segment observing periods**. Due to temporal heterogeneity of human behavior, telecommunication events are not captured uniformly over time. Therefore, it is common and effective to divide the data's collecting period into segments of study, *e.g.*, daytime and nighttime hours [15], weekdays and weekends [18], weeks or months [5]. Despite of possible loss of long-term behavior, this practice can usually ensure more users than using the whole collecting period.
- **Correlate with the mobility loss**. This is to build a function between results and inherent features of human footprints (*e.g.*, the location loss rate) by leveraging ground-truth datasets. The function is then used to fix the biased result obtained by the incomplete mobility information. For instance, Song *et al.* [5] find a linear correlation between the loss rate of voice calls and the logarithm of the entropy rate of time-ordered locations [5], and then employ this correlation to compute the predictability of human mobility from incomplete CDR-based mobility data.
- **Perform controlled experiments**. This is to repeat the methodology or the mobility feature computation on controlled datasets, *e.g.*, in [5, 16]. The controlled dataset usually has a higher resolution than the counterpart. The

[1] https://www.opencellid.org
[2] https://www.data.gouv.fr/fr/datasets/
[3] http://developers.google.com
[4] http://unwiredlabs.com
[5] http://opensignal.com
[6] https://location.services.mozilla.com

conclusion is more convincible provided that the same results can be obtained from both datasets.

- **Fill spatiotemporal gaps**. Although CDR cannot provide fully complete mobility information [15], it is enough to conduct reliable mobility inference so as to enlarge the availability of human footprints. Although the literature on this topic is fairly thin, several solid works are proposed. Ficek *et al.* [14] propose a probabilistic inter-call mobility model to determine users' positions between their consecutive voice calls. Sahar *et al.* [19] proposes an interpolation-based approach while it only work in the presence of trajectories composed of thousands of locations per day. For that, we also propose machine learning strategies to extend the availability of CDR having low user sampling rates [15, 20].

In summary, a solid data preliminary step is critical to conduct reliable human mobility analyses. To achieve such a step, the practices mentioned above need to be utilized in a comprehensive and flexible way corresponding to actual research or application scenarios.

IV. How individual mobility is measured?

Following the data preliminary step, the mobility of each user of the dataset is usually investigated as the next step by computing several straight-forward mobility features, to help the design and implementation of complex mobility analyses or applications. In this section, we first summarize these common features of individual mobility, in terms of users' *locations* and *travels*, and then we discuss an important but ignored issue of the computation of the radius of gyration.

A. How locations are visited?

In a CDR dataset, each user has a CDR-based trajectory of locations described by tens or hundreds of spatiotemporal points. It is essential to understand *how the user has visited these locations*. Several features are typical to answer this question, introduced as follows.

1) Cell coverage: Voronoi tessellations are often computed from all observed locations and are used as an estimation of the dataset's spatial resolution, as in [5, 15, 17]. Actually the locations of CDR are usually the ones of the cell towers handling telecommunication events. Mobile devices are actually in the areas covered by these cell towers. As an illustration, we plot in Fig. 2a the Voronoi tessellations of our voice call CDR dataset. We see that each Voronoi tessellation occupies an area around 2 km^2. In addition, we show in [15] that the location precision of using CDR dataset in metropolitan areas is around 1 km. Besides, with a large-scale dataset, such Voronoi tessellations can be leveraged to compute the population density of the area [2].

2) Repetitiveness: It is known that each user tends to have a few frequently visited locations [6, 16]. Therefore, it is important to understand the repetitiveness of these locations. Given a CDR-based trajectory with multiple locations, the repetitiveness on a per-user basis is computed as the number of unique locations in the trajectory and the probability of each location's appearance. We plot in Fig. 2b the overall probabilities $P(L)$ of appearance of the most frequent 50 locations versus their appearance rank L in the CDR dataset. We see that only two locations are visited more than 10% of time on average. Besides, it is observed that $P(L) \sim (L)^{-1}$ as in the other CDR datasets [16].

3) Significant locations and categories: It is also often seen in the literature to mark those frequently visited location with intuitive labels, *e.g.*, extracting important locations. For that, a simple and common way is to divide the observing period into sub-periods on a daily basis and to select the most frequent locations of each sub-period, such as *home* (nighttime) and *work* (daytime), as in [16, 18, 21].

B. How users travel?

The features above describe the mobility of a user from the viewpoint of locations. We also need to understand *how a user travels* during the observing period, from the viewpoint of his entire trajectory, which are usually described by the following features.

1) Displacement: The traveled distance between each two consecutive spatiotemporal points, *i.e.*, Δ_u, is computed to express the location displacement of a user, as in [16]. On a per-user basis, the maximum displacement $\Delta_u^{\max}$ and the average displacement $\overline{\Delta_u}$ often appears in the literature. We plot the distribution of the latter metric in Fig. 2c. It shows that a majority of the users (90%) have short-range movement (≤ 10 km) between two consecutive locations.

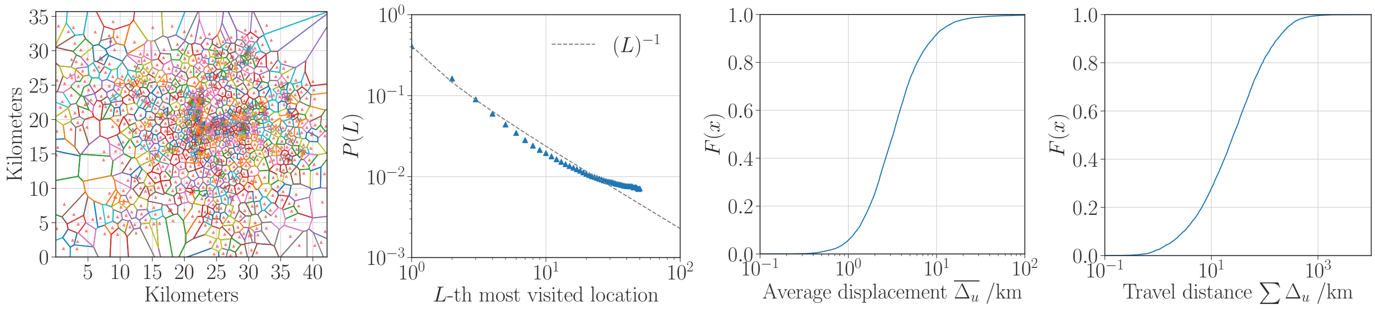
2) Traveled distance: The total traveled distance of a user, represented as $\sum \Delta_u$, is computed as the sum of a user's location displacements and show directly the user's movement, which is usually used with the radius of gyration together. We plot in Fig. 2d the distribution of $\sum \Delta_u$ across our users of study. We see that a large number of users have small traveled distances because of their low average location displacement and limited numbers of voice calls, while there is still a certain group of users who travel a lot. According to our experience, these users should be carefully addressed in the data preliminary.

3) Span of movement: To show how a user moves in a simple and quantitative manner, the radius of gyration of movement is often considered. After being originally adopted in human mobility in [16], the radius of gyration has become popular in human mobilities studies [10]. It is actually the perpendicular distance from the point mass to the axis of rotation, originally leveraged to deal with multi-dimensional points in structural engineering or polymer physics. For human mobility investigation, the radius of gyration is computed on a per-user basis from the locations of each trajectory. However, since the locations are spatiotemporal points in this case, how to deal with their temporal factors raises a novel and unresolved issue regarding the computation of the radius of gyration, particularly discussed next.

C. Computing radius of gyration from spatiotemporal points

When computing the radius of gyration from a CDR-based trajectory, we have to deal with the situation that a location is likely to appear many times. In other words, the spatiotemporal points of a trajectory may contain a far less number of unique cell tower locations, which raise a question: *how to regard such spatial repetitiveness in the radius of gyration?* Surprisingly, we find that mobility studies that compute the radius of gyration do not mention how they address this problem except a few (*e.g.*, [16]). Here we provide a thorough discussion regarding this issue.

The simplest way is to ignore temporal information and use only the *unique* locations in the trajectory. By doing this, we just consider those locations as normal points in a typical 2-dimensional space. Suppose a CDR-based trajectory has N



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

Fig. 2: (a) Voronoi tessellations in our area of study; red dots represent cell towers. (b) Probability of appearance of the most frequent 50 locations of each user; locations are ranked by their appearance frequencies on the x-axis. (c)(d) Cumulative distributions of each user's (c) average displacement and travel distance across our users of study.

unique locations $\{\mathbf{r}_1, \cdots, \mathbf{r}_N\}$, its corresponding radius of gyration, represented as RG_{unique}, is computed as follows:

$$RG_{\text{unique}} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\mathbf{r}_k - \mathbf{r}_{\text{cm}}^{\text{unique}})^2}, \;\mathbf{r}_{\text{cm}}^{\text{unique}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{r}_k, \tag{1}$$

where $\mathbf{r}_{\text{cm}}^{\text{unique}}$ is the center of mass of these unique locations. This computation avoids considering temporal dynamics of the user's movement and follows the general definition of the radius of gyration. Nevertheless, it cannot reflect the actual user's movement: the user's center of mass of $\mathbf{r}_{\text{cm}}^{\text{unique}}$ is strongly biased because those locations which the user stays a majority of time are regarded as equal as the occasional locations in Eq. (1).

The second way is to use the spatiotemporal points as they are and take all the points into account even if some of them are repeated, as used and described in [16]. It equals to use the locations' numbers of events (CDR) as their weights of importance in the radius of gyration. Suppose the locations $\{\mathbf{r}_1, \cdots, \mathbf{r}_N\}$ of the trajectory above have $\{m_1, \cdots, m_N\}$ events, respectively. The corresponding radius of gyration, represented as RG_{event}, is computed as follows:

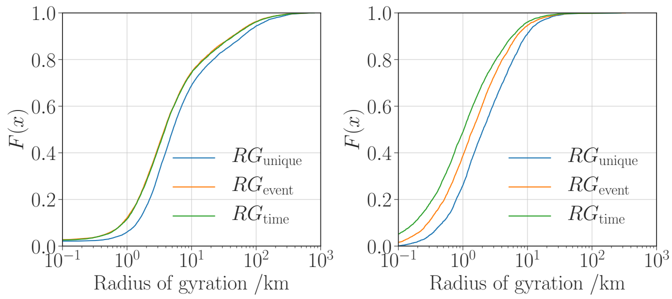
$$RG_{\text{event}} = \sqrt{\frac{\sum_{k=1}^{N} m_k(\mathbf{r}_k - \mathbf{r}_{\text{cm}}^{\text{event}})^2}{\sum_{i=1}^{N} m_i}}, \tag{2}$$

$$\mathbf{r}_{\text{cm}}^{\text{event}} = \frac{\sum_{k=1}^{N} m_k \mathbf{r}_k}{\sum_{i=1}^{N} m_i}. \tag{3}$$

For voice call CDR, this computation respects the user's movement because those locations with longer dwelling time usually have more voice calls [13] and higher importance in Eq. (2). However, it may be biased in internet visit CDR, the number of which is determined by not only dwelling time but also internet services and applications.

Therefore, we present the third and most reasonable way of the radius of gyration computation, *i.e.*, to divide the trajectory into time slots using a fixed temporal resolution and gather the most frequent location of each time slot. It can relax the impact of bursting events but can still extract the importance from the number of events. Accordingly, if the locations $\{\mathbf{r}_1, \cdots, \mathbf{r}_N\}$ occupy $\{s_1, \cdots, s_N\}$ time segments, respectively, the radius of gyration RG_{time} is computed as follows:

$$RG_{\text{time}} = \sqrt{\frac{\sum_{k=1}^{N} s_k(\mathbf{r}_k - \mathbf{r}_{\text{cm}}^{\text{time}})^2}{\sum_{i=1}^{N} s_i}}, \tag{4}$$



(a) (b)

Fig. 3: CDF of the three radius of gyration across the users of (a) the voice call CDR dataset and (b) the internet visit CDR dataset.

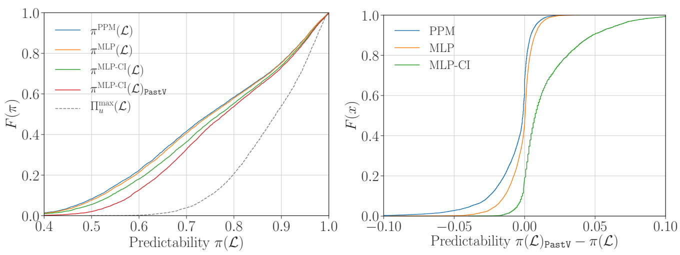
where the center of mass $\mathbf{r}_{\text{cm}}^{\text{time}}$ is computed similarly as in Eq. (3) by replacing all m_i with s_i.

To evaluate this three ways of computation, we compute them on both our voice call and internet visit CDR datasets, and portray in Fig. 3 the distributions of RG_{unique}, RG_{event}, and RG_{time} where RG_{time} is computed using 30-minute time slots. First, we observe that RG_{unique} is far larger than the other two metrics, indicating a strong bias brought by ignoring temporal factors. Then in the voice call CDR dataset, the distributions of RG_{time} and RG_{event} are quite similar, as shown in Fig. 3a. This is because the voice call CDR of a user is usually sparse in time and each 30-minute time segment tends to have only one or two calls so that the weights computed from time segments and events are highly similar. A large shift between these two distributions is observed in Fig. 3b, indicating that the burst of internet visits biases the radius of gyration if we still employ RG_{event}.

Consequently, to have realistic measurement of the user's movement span via the radius of gyration, whether or not CDR are sparse, we should measure each trajectory using an appropriate temporal resolution and adopted the time-segment-based metric as in Eq. (4).

V. Leveraging Individual Mobility at the Edge

Still, mobility data need to be leveraged by practical applications deployed at the edge of networks. This section presents our efforts on converting CDR-based mobility data into such applications by giving two representative applications that utilize the mobility of individuals as examples, *i.e.*, *mobility reconstruction* and *location prediction*.



(a) (b)

Fig. 4: (1) CDF of the theoretical and practical accuracy of forecasting a user's next cell tower from preceding ones. (2) CDF of the prediction accuracy enhancement by leveraging the knowledge of a user's preceding data traffic generation.

A. Mobility reconstruction

Due to the heterogeneous nature, voice call CDR do not have a stable sampling rate and cannot fully capture one's entire trajectory. For that, we address the mobility reconstruction problem to recover missing locations in a CDR-based trajectory, which is also valuable to trajectories obtained from other CDR types or mobility data proxies because the risk of loosing mobility information always exist. Nevertheless, the literature on this topic is fairly thin [15].

To fill the research gaps, we have designed the mobility reconstruction strategies using Gradient Boosting with decision trees [15] and Matrix/Tensor Factorization [20]. We also implement the state-of-the-art interpolation method for CDR-based trajectory reconstruction [19]. Leveraging our CDR datasets, we show the validity of these strategies via data-driven simulations. More importantly, these strategies only rely on an individual's trajectory and recent mobile devices with AI chips have enough computation power to implement these techniques. We believe that the mobility reconstruction is a reasonable application scenario deployed at the edge of cellular networks.

B. Location prediction

The accurate knowledge of a user's future whereabouts is significant in mobility-related applications, *e.g.*, optimizing energy consumption of mobile devices [22]. In our study, we consider relatively "simple" location prediction methods, to ensure that these methods can be implemented and meet the availability of computation and data storage on mobile devices. For example, simple Markov chain can achieve a fairly good accuracy in predicting a user's next location [23], and clearly, has low cost of time and space. Recent enhanced mobile devices, such as mobile AI chip integration [24] , make it possible to consider improved prediction methods. Particularly, we employ the following ones:

- **PPM** (Prediction by partial matching): a prediction method improved from Markov chain. It achieves better accuracy and requires less preceding samples [25].
- **MLP** (Multilayer perceptron): a classical machine learning method employing neuron networks [26]. For the feasibility of mobile phone deployment, we employ a simple full connected (256,256,256) network as inner layers and the rectified linear unit (ReLU) as the activation function. Named by the input context, we design three MLP-based predictors, *i.e.*, `MLP` – only using preceding locations, `MLP-CI` – using both preceding locations and temporal (weekday, date, hour) features, and `MLP-CI-PastV` – adding features of mobile data traffic consumption into `MLP-CI`.

With the help of CDR datasets, we can study the location prediction problem and enlarge the population scale to thousands of users. Particularly, we preform our study on approximately 7K users with sufficient mobility data in an observing period of 150 days. For each predictor and each user, we let the predictor initialize using the locations of the first 100 days to guarantee a entire "warm up", and using it to predict the remaining locations and compute the accuracy. During the prediction, each predictor is updated every day to simulate an actual mobile phone application.

We evaluate the performance of our predictors and portray the CDF of the prediction accuracy across our users in Fig. 4a, π^{PPM}, π^{MLP}, $\pi^{\text{MLP-CI}}$, and $\pi^{\text{MLP-CI}}_{\text{PastV}}$ show the actual prediction accuracy of each user, and $\Pi_u^{\max}$ represents the theoretical performance derived via information theory. $\Pi_u^{\max}$ worths some additional explanation. It is computed via information theory [5] and shows the upper bound of the prediction accuracy from temporal orders of historical visiting patterns. We see that, the theoretical upper bound shows an 85% of the maximum expected accuracy on average, while leveraging preceding locations can only achieve 73% (`PPM`) and 74% (`MLP`) of the average practical accuracy. Approximately 76% of the average practical predictability is achieved by the `MLP-CI` predictor which further leverages the time as the context information. The best performance is achieved by the `MLP-CI` predictor with the knowledge of previous data traffic volumes, which has 79% of the practical predictability.

We also plot in Fig. 4b the CDF of the accuracy enhancement of each user brought by the use of historical data traffic volumes in the prediction. We note that for the `PPM` and `MLP` predictors, only less than 50% of the users have such enhancement up to 5%, while the practical predictability of the results even describes at most 10% surprisingly. It indicates that, the context information, such as time and data traffic consumption, do have the capability of achieving a better prediction of a user's locations, while only the machine learning techniques could absorb and utilize such information efficiently, nor the Markovian methods.

In summary, we find that simple prediction methods can achieve a fairly good accuracy in location prediction and are able to be deployed in mobile phones.

VI. Conclusion and Discussion

In the previews sections, we have presented the important issues in terms of data collection, data preliminary, mobility feature completion, and mobility applications. Still, because CDR have appeared as important resources for research since only the past decade [2], there are a multitude of remaining open problems and future research directions regarding human mobility. In the following, we discuss some critical issues.

- **Is there any better mobility data source?** This answer to this question depends on application scenarios. For instance, GPS data is usually a better choice providing higher spatiotemporal resolutions, if a large-scale user population is not necessary. There is general agreement on the fact that no other technique can cover the same amount of users as CDR and meanwhile maintain such low cost. In fact, CDR data is still far from its full potential as mobility data source. With increased positioning techniques and enough CDR types released, CDR can keep almost the same spatiotemporal granularities.

Obtaining such data requires cellular operators with better openness and addressing non-technical issues such as privacy and security.

- **Can mobility reconstruction models perform better?** Inferring missing mobility data from CDR captured is quite a useful data preliminary practice and does not receive enough attention, as discussed in Section III. The current relevant techniques, including ours, mainly utilize the repetitive human mobility patterns. Mobile information can be extracted from CDR and may contribute to mobility reconstruction. For instance, with multiple CDR types, one can reasonably expect having coarse-grained long-term mobility information of users and finer-grained short-term mobility information of the same users only in some partial observing periods. No existing work studies how to assess the long-term mobility reconstruction problem using such mixed information. Besides, recovering a user's trajectories may benefit from knowing similar trajectories of other users.
- **How to improve human mobility predicative models?** Forecasting future human whereabouts is one of the most important topics of human mobility investigation [10, 11]. So far, relevant studies have covered a variety of techniques such as Markov chains, time series analysis, Naive Bayes, Nonparametric Bayesian inference, and even artificial neural network, considered from single-user models to aggregated models, and analyzed both theoretical and practical predictability of individual mobility. However, there is still a research direction that is nearly untouched, *i.e.*, leveraging contextual information into mobility prediction. For instance, when working with locations, mobile network traffic (*e.g.*, data traffic as in Section V-B) also described by CDR and can contribute to mobility prediction. We believe that more context data (*e.g.*, points of internet, web browsing, and environment of mobile devices) have such power to be discovered. Moreover, as collecting such data requires deeper mobile device integration and collaboration, there is a huge space of possible fog computing applications.

Consequently, in this paper we surveyed the literature on utilizing CDR into human mobility studies, and provided the major takeaways in terms of CDR data mining, along with open research directions.

REFERENCES

[1] P. Garcia Lopez, A. Montresor, D. Epema, A. Datta, T. Higashino, A. Iamnitchi, M. Barcellos, P. Felber, and E. Riviere, "Edge-centric computing: Vision and challenges," *SIGCOMM Computer Communication Review*, vol. 45, pp. 37–42, Sept. 2015.

[2] 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.

[3] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. L. Barabasi, "Human mobility, social ties, and link prediction," in *SIGKDD 2011*, pp. 1100–1108, ACM, 2011.

[4] G. Chen, S. Hoteit, A. C. Viana, M. Fiore, and C. Sarraute, "The spatiotemporal interplay of regularity and randomness in cellular data traffic," in *LCN 2017*, pp. 187–190, IEEE, Oct 2017.

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

[6] E. R. O. Mucelli, A. C. Viana, C. Sarraute, J. Brea, and I. Alvarez-Hamelin, "On the regularity of human mobility," *Pervasive and Mobile Computing*, vol. 33, pp. 73–90, Dec. 2016.

[7] R. K. Barik, H. Dubey, A. B. Samaddar, R. D. Gupta, and P. K. Ray, "Foggis: Fog computing for geospatial big data analytics," in *UPCON 2016*, pp. 613–618, IEEE, 2016.

[8] H. Dubey, J. Yang, N. Constant, A. M. Amiri, Q. Yang, and K. Makodiya, "Fog data: Enhancing telehealth big data through fog computing," in *ASE BigData & SocialInformatics 2015*, p. 14, ACM, 2015.

[9] 3GPP, "Telecommunication management; Charging management; Charging Data Record (CDR) parameter description (15.2.0)," TS 32.298, 3GPP, Mar. 2018.

[10] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Physics Reports*, 2018.

[11] E. Toch, B. Lerner, E. Ben-Zion, and I. Ben-Gal, "Analyzing large-scale human mobility data: a survey of machine learning methods and applications," *Knowledge and Information Systems*, pp. 1–23, 2018.

[12] Y. Dong, J. Tang, T. Lou, B. Wu, and N. V. Chawla, "How long will she call me? distribution, social theory and duration prediction," in *ECML PKDD 2013*, pp. 16–31, Springer, 2013.

[13] G. Ranjan, H. Zang, Z.-L. Zhang, and J. Bolot, "Are call detail records biased for sampling human mobility?," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 16, p. 33, Dec. 2012.

[14] M. Ficek and L. Kencl, "Inter-call mobility model: A spatio-temporal refinement of call data records using a Gaussian mixture model," in *2012 Proceedings IEEE INFOCOM*, pp. 469–477, IEEE, Mar. 2012.

[15] G. Chen, S. Hoteit, A. C. Viana, M. Fiore, and C. Sarraute, "Enriching sparse mobility information in call detail records," *Computer Communications*, vol. 122, pp. 44 – 58, 2018.

[16] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, June 2008.

[17] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, Mar. 2013.

[18] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Lecture Notes in Computer Science*, pp. 133–151, Springer, 2011.

[19] 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.

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

[21] P. Baumann, W. Kleiminger, and S. Santini, "How long are you staying?: predicting residence time from human mobility traces," in *MobiCom 2013*, pp. 231–234, ACM, 2013.

[22] Y. Chon, E. Talipov, H. Shin, and H. Cha, "Mobility prediction-based smartphone energy optimization for everyday location monitoring," in *SenSys 2011*, pp. 82–95, ACM, 2011.

[23] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific reports*, vol. 3, p. srep02923, 2013.

[24] S. Guo, V. Leung, and X. Yao, "Guest editors introduction: Intelligence in the cloud," *IEEE Cloud Computing*, vol. 4, no. 6, pp. 34–36, 2017.

[25] A. Moffat, "Implementing the PPM data compression scheme," *IEEE Transactions on communications*, vol. 38, no. 11, pp. 1917–1921, 1990.

[26] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.