Predictability and Prediction of Human Mobility Based on Application-collected Location Data

Sihan Zeng*, Huandong Wang*, Yong Li*, Depeng Jin*
* Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
Email: liyong07@tsinghua.edu.cn

***Abstract*—In the modern information society, accurate prediction of human mobility becomes increasingly essential in various areas such as city planning and resource management. With users' historical trajectories, the inherent patterns of their movements can be extracted and utilized to accurately predict the future movements. In this paper, based on a dataset of 100,000 individuals' actively uploaded location information collected by apps, we discover the average theoretical limits of the predictability to be as high as 93%. Since the app-collected data contains the physical context of the location, we implement a clustering method based on the contextual information that cluster the locations into three divisions, street, district and region. In order to solve the unevenly distribution and the high missing rate of the application collected location data, we firstly use the Gibbs sampling algorithm to complete the missing data of the trajectory and then employ a high-order Markov chain model to predict the most likely locations visited by each user. Result shows that our prediction algorithm can achieve accuracy as high as 67%, 78%, 87% for the three context-based divisions respectively, which are 10% higher on average than the divisions without context. In addition, the correlation coefficient between prediction accuracy and predictability reaches as high as 0.86. Finally, we investigate various factors including spatial and temporal resolution, orders of Markov models, radius of gyration, in order to explore the predictability under different circumstances.**

I. Introduction

With the rapid development of the wireless and networking technology, mobile neworks have imposed a profound impact on people's daily life for their marvelous capability. These applications utilize the users' current and historical location information records (LIR) to analyze their mobility patterns to enable numerous applications, such as targeted advertising, city planning and smart navigation.

Generally speaking, the LIR data collected from the mobile networks can be divided into two categories, i.e., data collected by Internet service provider (referred to as ISP-collected data) and data collected by applications (referred to as app-collected data). The ISP-collected data are passively and periodically collected regardless of behaviors of the users. This sort of data preserves the complete and consecutive trajectory of each user. Most of the existing studies are based on users' ISP-collected location data [1-7]. Song *et al.* [1] quantified the predictability in human mobility by studying the regularities shown in the trajectory. According to their studies, the potential predictability reaches 93% on a mobile phone record dataset. Wang *et al.* [2] link the human mobility with the social network, by segregating the similar users using the information from social media, more general and universal mobility patterns on a certain group of people were extracted, suggesting the huge predictability of the individual's movements. Moreover, applying the predictability into practice, many researches have also been conducted on the prediction of human mobility on various models, such as Markov Chain models [3], [4], neural network [5], Bayesian network [6], finite state machine [7]. On the other hand, however, few researches have focused on either predictability or prediction algorithm on app-collected location data. This aspects of researches remain to be explored.

In comparison to ISP-collected data, app-collected LIR data is actively triggered by users themselves in applications. This kind of location data will be collected when using the applications while the location information of the rest time remains unknown. It is exactly the characteristics of the app-collected data that arouses several difficulties to our study. First, the app-collected data contains the physical context of the location because the purpose of using the application certainly correlates with the location recorded, e.g., ordering a taxi, searching a restaurant. Such correlations provide valuable information to analyze the human mobility patterns. However, simple grid for the city apparently lose the information. Hence, it is essential to find a proper spatial division of the city to reserve the physical context of app-collected data. Second, the app-collected data are partially missing since usually the applications do not recordm users' locations when they are not using the apps. Third, the app-collected data are heterogeneous in spatial and temporal domain since the time when people use the application is unevenly distributed. Under these circumstances, the methods aroused in the previous study apparently are not suitable for accurate predictions on the dataset. We need to propose new methods to adapt to these features of the app-collected data.

In this paper, we address the above three challenges to facilitate the analysis. Our work can be summarized as follows:

- In order to reserve the physical context of the locations, we contextually cluster the locations into multiple non-overlapping districts of the city instead of using fixed coordinate grid that will lose the physical context. We also compare the predictability and the prediction accuracy between the two divisions to analyze the effect of context on prediction. Results reveal that the trajectories on context-based division are more predictable than those on division without context under the same spatial granularity.
- We design a Markov-based method using Gibbs sampling

2155-6814/17 $31.00 © 2017 IEEE
DOI 10.1109/MASS.2017.32

to solve the unevenly distribution and the high missing rate of the app-collected data. By restoring the trajectory, we estimate the transition matrix to make prediction of users' movement. Results show that, based on app-collected dataset, our method achieves the same accuracy of the previous studies on the ISP-collected dataset.
- In order to investigate the effect of heterogeneity, we carry out a thorough analysis of the predictability and prediction accuracy based on our designed method on the app-collected dataset. The varying factors include the spatial and temporal resolution, the orders of Markov models, the radius of gyration etc.

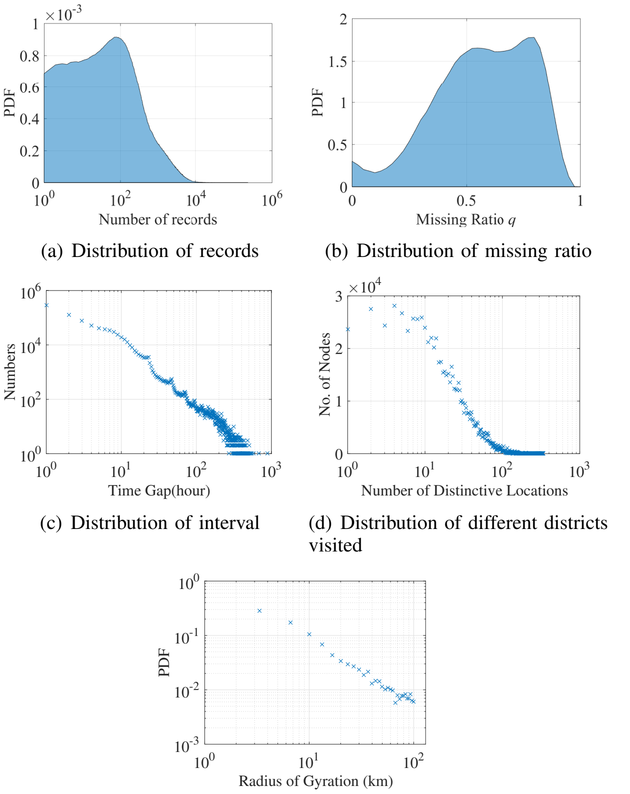
The rest of this paper is organized as follows: Section II describes the dataset. In section III, we introduce the methodology and metrics. Experiments and results will be presented in Section IV. Then, we discuss the related works in Section V and conclude the paper in Section VI.

II. Dataset

Our dataset was collected by a popular localization platform. As mentioned in introduction, when users use related Apps, such as WeChat (most pervasively used online instant messenger in China), their location records (LIR) will be uploaded to the servers and collected by this platform. Thus, these records strongly imply users' behavior patterns. Overall, the analysis of this paper is based on a dataset collected on LIR from 1,000,000 anonymous users, who are active during September 17, 2016 to October 31, 2016 in Beijing. There are 800 million records in our data set, which is a large-scale dataset and guarantees the credibility of our study. Each entry consists of the following fields: the anonymized ID of the user, the time of the record (accurate to second), and the location information in the format of GPS coordinates.

To further illustrate the characteristic of app-collected data, we present the statistic metrics as follows. Probability density function of the number of each users LIR is shown in the Fig. 1(a). Unlike the log-normal distribution of ISP-collected data [3], large proportion of the user in our dataset have relatively small amount of records, which exactly indicates that our dataset is highly sparse and heterogeneous. If we define the proportion of the missing data of a user to be parameter q, as a demonstration of the incompleteness of a user's location information. The distribution of q is presented in Fig. 1(b). Large part of the users have missing ratio q more than 0.5, suggesting again that app-collected data is partially missing. In terms of time, we count the intervals of consecutive two records, and show its distribution in the Fig. 1(c). The character of unevenly distributed is obviously showing, with average interval of consecutive two records achieve as high as 3.11 hours.

On the other hand, in terms of spatial aspect, we also present several findings concerning our app-collected dataset. Distribution of the total number of different districts visited by each users is shown in Fig. 1(d), with 72% of users recording in less than 30 locations during the 46 days, and few recoding in more than 100. The distribution of the radius of gyration, r_g of sampled users is shown in Fig. 1(e), demonstrating a significant decay as r_g increases, which indicates that the movements of the majority of users were confined within an area of 30km.



(a) Distribution of records

(b) Distribution of missing ratio

(c) Distribution of interval

(d) Distribution of different districts visited

(e) Distribution of radius of gyration

Fig. 1: **Characteristics of app-collected dataset**

In general, our dataset has highly sparse and heterogeneous features in the temporal aspect, yet it shows huge regularity of users' mobility patterns in the spatial aspect. Hence, if given the proper method, the users' movements of app-collected data can be as predictable as those of ISP-collected data, which will be further discussed in the following section.

III. Method

In this section, we first introduce the process of trajectory building together with the method of calculating entropy and limits of predictability based on trajectories. Then, we present our proposed prediction method. For readability, we summarize all the notations used in this paper in Table I.

A. Trajectory Building

To investigate the predictability of users' mobility, firstly we need to build users' trajectories based on the LIR by determining the time interval and spatial division. In order to accurately reveal the mobility patterns of each user, while restraining the computational complexity, we carry out the subsequent studies on the time interval of 1 hour, 2 hour and

TABLE I: List of commonly used notations.

Notation	*Description*
T	The original trajectory.
R	The restored trajectory.
$\mathscr{R}$	The set of all possible restored trajectories.
S	The entropy of the trajectory.
Π	The maximum predictability of the trajectory
i	A state in the state space of the Markov model.
M	Total number of distinctive states.
L	The length of the trajectory.
L'	The total number of missing points of the trajectory.
s_n	A subsequence with length n.
P	Transition matrix.
P_{s_n}	The row of P indexed by s_n.
$P_{s_n i}$	The ith entry of row P_{s_n}
γ	The prediction accuracy.
$\sigma(q)$	A parameter used in estimation of entropy.
Λ_i	The length of the shortest substring starting at position i which does not previously appear from position 1 to i.

4 hour. We add one data point of location to a user's trajectory every time interval. If a user visits more than one locations within one interval, his most visited place will be marked as location of this time interval in his trajectory. In comparison, if there does not exist an LIR for a user during certain time interval, that point of the trajectory remains unknown. Specifically, we let $T = (X_1, X_2, ..., X_L)$ denote the sequence of locations where a user was observed at each time interval, in which L is the length of the sequence. For the unknown time interval k, we let X_k to be uncertainty.

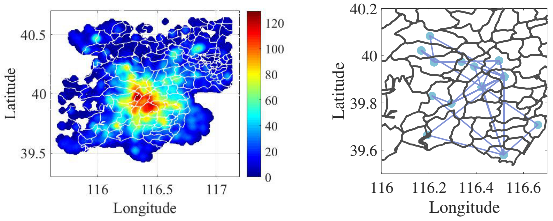
On the other hand, as for spatial division scale, in order to reserve the physical context of the locations, we contextually divide the city into multiple non-overlapping regions of three different sizes, denoted separately by street, district and region. Correspondingly, Beijing has 188 regions, 6353 districts and 65233 streets, respectively. To be specific, for the division of street, we segregate the locations with similar functions on a street level, since the sites of similar Points of Interest (POI) usually gather in the same area. Hence, we can regard the a segmented street as a unit that carries the context of certain category, such as residence and entertainment areas, which is exactly the physical context [8] . In addition, we cluster these streets into larger segments, i.e., districts and regions, for further study. Compared with the geographic grid, a street is of the same scale as a 200m grid and a region is the same scale as a 4km, whereas these grids provide no physical context of the locations, as summarized in Table II. Hence, the three context-based divisions are chosen for trajectory building. An overview of geographical locations density with the division of region is shown in the Fig. 2(a). A typical trajectory of under division of street is depicted in Fig. 2(b).

B. Entropy and Predictability

We choose one of the most fundamental quantities capturing the degree of predictability, i.e., the entropy to evaluate the regularity of a trajectory, which is defined as the expected value of the information contained in each element of the sequence. For instance, if a trajectory has the entropy of S, then every location of the trajectory contains S bits of new information on average, which means the user will choose his/her next move in 2^S locations randomly. Specifically, real entropy is given by $S^{\mathrm{real}} \equiv -\sum_{T_i \in T} p(T_i) \log_2 p(T_i)$, where T_i is a subsequence of the trajectory and T represents the set of all T_i and $p(T_i)$ is the probability that T_i is found in T [1]. The real entropy considers both temporal and spatial correlated factors. Therefore, it can provide the best evaluation of the regularity of the trajectory. In addition, we introduce another two metrics of the entropy to facilitate the subsequent analysis, which are defined as follow:

TABLE II: Features of different spatial divisions

Spatial division	*Resolution*	*Physical Context*
Street	High	√
200m grid	High	×
Region	Low	√
4km grid	Low	×



(a) Geographical locations density (b) Typical trajectory

Fig. 2: **Overview of spatial division on app-collected data**

- Random entropy given by $S^{\mathrm{rand}} \equiv \log_2 N$, where N is the number of distinct locations visited by the user. This measurement illustrates the regularity in the precondition that the user visits each location with equal probability.
- Temporal-uncorrelated entropy given by $S^{\mathrm{unc}} \equiv -\sum_{j=1}^{N} p(j) \log_2 p(j)$, where N has the same meaning as above while $p(j)$ represents the probability that location j is visited by the user. This measurement takes the frequency of different locations visited by users into account, yet ignoring the visiting sequence.

Naturally, the three forms of entropy for the same user's trajectory follow the inequality: $S^{\mathrm{real}} \leq S^{\mathrm{unc}} \leq S^{\mathrm{rand}}$.

From the expression of each entropy, we can see that both S^{rand} and S^{unc} can be easily derived from the trajectory, while the S^{real} requires much more complicated calculations, of which the computational complexity reaches $O(N^4)$. To calculate the real entropy with high accuracy and efficiency, we use an estimator based on Lempel-Ziv data compression [9], the real entropy of a trajectory with n points can be estimated by

$$S = \left(\frac{1}{n} \sum_i \Lambda_i \right)^{-1} \ln n, \tag{1}$$

where Λ_i is the length of the shortest subsequence starting at position i which does not previously appear from position 1 to

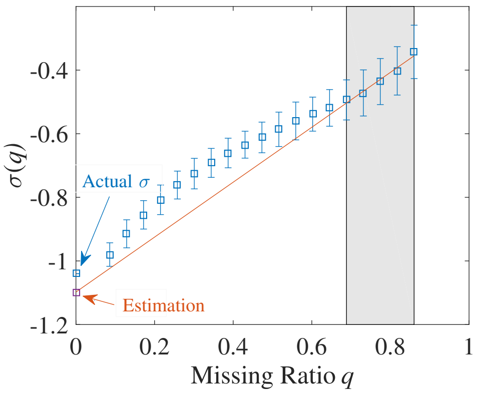


Fig. 3: **Linear relationship between** ($\sigma(q)$ **and** q)

$i-1$. Kontoyiannis *et al.* [9] have proved that the expression above converges to the real entropy when n approaches infinity.

Apparently, the above measurements of entropy only apply to the complete trajectories. Once the missing ratio q increases, the entropy $S(q)$ will accordingly increase since the missing data introduces extra uncertainty. According to the research of Song *et al.* [1], the parameter $\sigma(q) \equiv \ln(S(q)/S^{\text{unc}}(q))$ is observed to have a linear relationship with q as shown in Fig. 3. Taking advantage of the property, we manually mask the known data to change the missing ratio q, at the same time calculate the parameter $\sigma(q)$. After obtaining several 2-tuples $(\sigma(q), q)$, we extrapolate the $\sigma(q)$ to $q = 0$ to estimate σ_{est} for a complete trajectory. The real entropy is then calculated as $S^{\text{est}} = e^{\sigma_{est}} S^{\text{unc}}$. The detailed process of calculating S^{est} is shown in the Algorithm 1.

Before predicting on the dataset, there is still one particular question we have to figure out in advance, which is to what degree can we make prediction about human mobility. In other words, we need to determine the maximum probability that we can make the accurate prediction in advance of designing an algorithm. We denote that maximum probability as Π . According to the studies of Fano *et al.* [10] and Navet *et al.* [11] , Π is subjected to the Fano's inequality $\Pi \leq \Pi^{\max}$. Specifically, if a user with entropy S moves between N locations, then the limit of predictability $\Pi^{\max}$ satisfies:

Algorithm 1 Real entropy estimation algorithm

```
Input: Trajectory $T = \{X_1, X_2, ..., X_L\}$
Output: Entropy $S^{\text{est}}$
procedure ENTROPY ESTIMATION($T$)
    Initiate $\mathscr{D} \leftarrow \emptyset$
    Calculate missing ratio $q$
    while $q < 0.9$ do
        $S \leftarrow$Randomly choose 5% of known data from $T$
        for all points $X_i \in S$ do
            Set $X_i \leftarrow$ ? to unknown
        end for
        $q \leftarrow q + 0.05$
        Calculate $S^{\text{unc}}(q)$ and $S^{\text{real}}(q)$ based on (1)
        $\sigma \leftarrow \ln(S^{\text{real}}(q)/S^{\text{unc}}(q))$
        $\mathscr{D} \leftarrow \mathscr{D} \cup (q, \sigma)$
    end while
    $\hat{\sigma}(q) \leftarrow$Linear regression on $\mathscr{D}(q, \sigma)$
    $S^{\text{est}} \leftarrow e^{\hat{\sigma}(0)} S^{\text{unc}}$
end procedure
```

$$S = H(\Pi^{\max}) + (1 - \Pi^{\max})\log_2(N-1), \tag{2}$$

where the binary entropy function $H(\Pi)$ is:

$$H(\Pi) = -\Pi\log_2(\Pi) - (1-\Pi)\log_2(1-\Pi), \tag{3}$$

In the previous part, we discussed how to derive the entropy of a trajectory. Therefore, we can easily calculate the limit of predictability by solving Equation (2). Like the three measurements of entropy, the limit of predicability also has three different forms – Π^{rand}, Π^{unc}, and Π^{real}. According to the inequality of Entropy S, we have:

$$\Pi^{\text{real}} \geq \Pi^{\text{unc}} \geq \Pi^{\text{rand}}, \tag{4}$$

In this paper, we mainly analyze the predictability based on Π^{real} since it best describes the limit taking both factors above in to account.

C. Prediction

Having known the limit of predicability, we then design our prediction algorithm to approach this limit. Specifically, we propose a prediction algorithm based on Markov model while using Gibbs sampling method to solve the problem caused by missing data.

In a Markov chain model, the movements of human are modeled as transitions among definite and countable states space, which can be denoted as region R with M states. If we assume the Markov chain to be time-homogeneous, then for each user there exists a unique transition matrix P of size $M \times M$, where M is the total number of the states. The entry P_{ij} represents the probability that the user moves from location i to location j. As for the n-order Markov chain, the order n means n previous locations are needed to derive the transition probabilities of the user transits from current state i to other state j in the next time slot. In other words, the next location of the user depends on the past n visited locations, under the circumstances, the transition matrix becomes a (n+1)-dimensions matrix. The entry $P_{i_0 i_1 \ldots i_n}$ represents the probability that the user move to location i_n given the historical n locations to be $i_0, i_1, ..., i_{n-1}$.

Because the n-order Markov chain is more complicated than homogeneous Markov chain, we define the row of (n+1)-dimension transition matrix P as the probability distribution of the user's next location giving his historical n locations. Hence, the rows of P that are indexed by n parameters, denoted as $P_{i_0 i_1 \ldots i_{n-1}}$. We use s_n denoting the subsequence $i_0, i_1 \ldots i_{n-1}$ of length n, then the rows of P becomes P_{s_n} and the i_{th} entry of the row becomes $P_{s_n i}$, which represents the probability of a user moving to location i with his historical n locations to be s_n.

After obtaining both transition matrix P and historical trajectory s of length n, the prediction of next location i^{est} becomes finding the maximum of the row of transition matrix P_{s_n}, which can be expressed as follow:

$$i^{est} = \arg\max_i P_{s_n i}. \tag{5}$$

Therefore, it is our primary task to estimate the transition matrix P based on the incomplete trajectories. We let R denote the restored trajectory of the part-missing trajectory T, then the estimation of transition matrix P^{est} is:

$$P^{est} = E[P|T]. \tag{6}$$

The distribution of P under the condition of T can be calculated by enumerating $R \in \mathscr{R}$, the set of all possible R . The procedure can be expressed as follow:

$$Pr(P|T) = \sum_{R \in \mathscr{R}} Pr(P|R,T). \tag{7}$$

Apparently, enumerating all the possible restored trajectory involves tremendous calculating work, with the complexity exponentially increasing as the length of trajectory increase. Hence, we employ Gibbs sampling, a Monte Carlo method that update both restored trajectory R and transition matrix P alternatively at the same time. First, we fill in the missing points with randomly chosen data from the locations the user ever visited. Then, the iterative algorithm can be expressed as follow:

$$P^{\{n\}} \sim Pr[P|R^{\{n-1\}}, T], \tag{8}$$

$$R^{\{n\}} \sim Pr[R|P^{\{n\}}, T]. \tag{9}$$

where (8) represents sampling transition matrix $P^{\{n\}}$ from the restored trajectory $R^{\{n-1\}}$ and the original one T. We assume that the rows of transition matrix are independent. By considering a Dirichlet prior for each row, we sample each row of the transition matrix P_{s_n} from the following distribution:

$$Dirichlet(\{Count_{s_n i}(R^{\{n-1\}}) + \varepsilon_{i_{n-1} i}\}_{i=1...M}). \tag{10}$$

where $Count_{s_n i}(\cdot)$ is the number of sequence $s_n i$ appearing in R. And $\varepsilon_{i_{n-1} i}$ is a small positive number to ensure the all the states have the probability to be sampled, even if the chances are small. Moreover, if there exists a physical barrier between i_{n-1} and i, $\varepsilon_{i_{n-1} i}$ shall be set to zero to prevent the case from sampling.

On the other hand, (9) represents sampling trajectory $R^{\{n\}}$ from the transition matrix $P^{\{n\}}$. We update the originally unknown data with the posterior probability distribution derived from transition matrix $P^{\{n\}}$ given the past n locations. For example, if we assume the order of the model is 2, and there is an originally unknown data in the lth of a trajectory T, with trajectory from $(l-2)_{th}$ to $(l-1)_{th}$ denoted as $R[l-2:l-1]$. Then, the probability of the $T[l]$ to be state i is:

$$P_{R[l-2:l-1]i} P_{R[l-1]iR[l+1]} P_{iR[l+1:l+2]}. \tag{11}$$

After normalizing the M probabilities as above (since there is M states), the distribution is obtained as follow:

$$\left\{ \frac{P_{R[l-2:l-1]i} P_{R[l-1]iR[l+1]} P_{iR[l+1:l+2]}}{\sum_{j=1}^{M} P_{R[l-2:l-1]j} P_{R[l-1]jR[l+1]} P_{jR[l+1:l+2]}} \right\}_{i=1,2...M}. \tag{12}$$

Thus, applying the distribution to the model with order n, for every originally missing points $m_l (l = 1, 2..L')$ with L' representing the total number of the missing points, we sample the location based on the following distribution.

$$\left\{ \frac{\prod_{k=0}^{n} P_{R^{\{i\}}[l-n+k:l+k]}}{\sum_{j=1}^{M} \prod_{k=0}^{n} P_{R^{\{j\}}[l-n+k:l+k]}} \right\}_{i=1,2...M}, \tag{13}$$

where $T^{\{j\}}$ represents the trajectory with missing data points m_l replaced by location r_j. The detail process is shown in the Algorithm 2.

Studies from Robert [12], [13] have proved the convergence of the Gibbs sampling for this problem, which means the estimated transition matrix P will eventually converge to the actual matrix that describes the movement patterns of the user. Meanwhile, the missing part of the trajectory can be fully restored.

Algorithm 2 Prediction Algorithm based on Gibbs sampling

Input: Trajectory $T = \{X_1, X_2, ..., X_L\}$, Threshold of iteration δ_{th}

Output: Restored Trajectory R, Transition matrix P

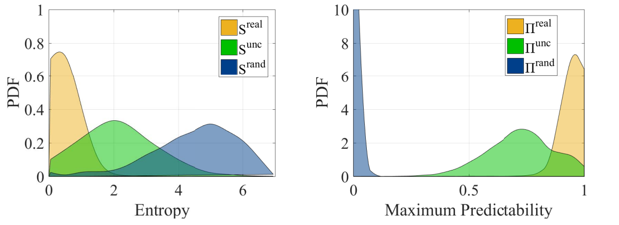
```
procedure PREDICTION($T$)
    Initiate $R$, $\delta$
    Set Threshold
    while $\delta > \delta_{th}$ do
        for all sequence $s \in$State Space do
            Count the number of $s$ appearing in $R$
        end for
        Sample every row of $P$ based on Equation (10)
        for all missing points $X \in R$ do
            Update $X$ from $P$ based on Equation (13)
        end for
        $\delta \leftarrow$ number of altered points of $R$
    end while
end procedure
```

IV. Performance Evaluation

In this section, we will first study the entropy and the limits of predictability. Then we make predictions by utilizing the introduced methodologies with the purpose to discover the moving patterns in human mobility.

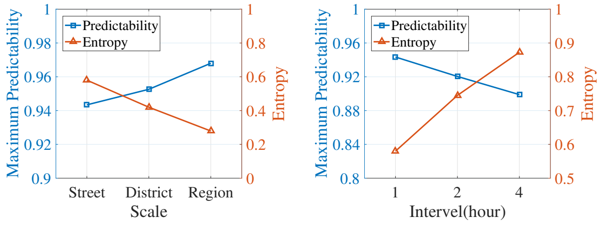
A. Entropy and Predictability

In the previous section, we introduce three metrics of the entropy on a trajectory, which is S^{rand}, S^{unc} and S^{real}. The former two can be directly derived from the trajectory, while the latter kind will be calculated by the estimation algorithm



(a) PDF of Entropy (b) PDF of Maximum Predictability

Fig. 4: **Distributions of entropy and predictability of three different measurements**



(a) Spatial Resolution (b) Temporal Resolution

Fig. 5: **Distributions of entropy and predictability under different spatial and temporal resolution**

introduced above. Since overmuch missing data will bring in considerably uncertainty, to avoid these interference while test our designed method, we choose the users with missing ratios q under 0.8 in the further study. Fig. 4(a) shows the distributions of three forms of entropy of the trajectory we build under street scale.

From the figure, three separated peaks of these distribution can be clearly observed, which represent S^{real}, S^{unc} and S^{rand}, respectively. The random entropy has relatively largest value with its distribution $P(S^{\mathrm{rand}})$ peaks at $S^{\mathrm{rand}} \approx 3.8$, which suggested that a typical user can be found randomly in $2^{S^{\mathrm{rand}}} \approx 14$ locations. This entropy provides an approximation of the total numbers of the locations visited by the user, while more factors will be considered in the other two forms.

The temporal-uncorrelated entropy S^{unc} lies in the middle and the real entropy of the trajectory has the smallest value. This is an intuitive result since former one considers the spatial factor and the latter one considers both spatial and temporal correlations, which are the very factors that introduce the regularity and eliminate the uncertainty. Specifically, S^{unc} peaking at 1.5 and S^{real} peaking at 0.8 indicate that if we fully explore the regularity of a trajectory, then the user's next movement can be predicted within 2 locations.

Then, we examine the limit of predictability utilizing the three entropy, Π^{rand}, Π^{unc} and Π^{real} based on Equation (2). The result is shown in the Fig. 4(b). As we expected, the decrease in uncertainty results in the increase of maximum predictability. If only given the numbers of the location that a user ever visited, there is little chance to accurately predict his next location, $\Pi^{\mathrm{rand}} \approx 0$. Furthermore, if we take the spatial factors into account, the maximum predictability Π^{unc} increases to a median value of 0.8, indicating the enormous regularity hidden behind the visiting frequency of different locations. Finally, if we consider both spatial and temporal correlations, the maximum predictability can reach as high as 0.94, an unexpectedly high level that may exactly illustrate the regularity regardless of the seemingly random human moving pattern.

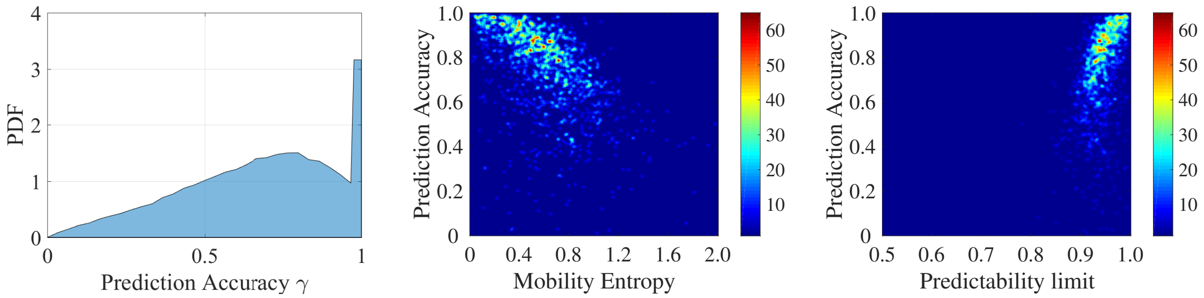
Also, we analyze the entropy and maximum predictability under different spatial and temporal resolutions. As discussed in the previous section, we introduce three scales of city division: street, district and region. The average entropy and maximum predictability of the three divisions are separately shown in Fig. 5(a). By comparing the result between different scales, we can concluded that trajectories on the large scale are more predictable with the smaller entropy. This is for an obvious reason that plenty of the users have their trajectories appeared in only one or two regions since the large scale divides Beijing into only 188 regions. Consequently, the uncertainty of their movements are of small level and the prediction on these users is relatively simple. In contrary, under the small scale, the entropy becomes larger and the predictability becomes smaller,indicating that uncertainty of the trajectory increases. The predictability's decreasing as the spatial resolution becomes higher is another intuitive result because a user will visit more locations in the smaller scale, which undoubtedly increases the difficulty to accurately predict his/her movement.

As for the temporal resolutions, we change the time interval of a trajectory from 1 hour to 2 hours and 4 hours. The average entropy and maximum predictability under the three temporal resolutions are shown in the Fig. 5(b). Different from the results of spatial resolution, as the time interval extended, the maximum predictability decreases instead. A potential explanation for the correlation is that the extended interval leads to the reduction of points for every trajectory. Only one point is sampled from two or four points of the original trajectory, which undoubtedly masks the information of users' movements and breaks the regularity hidden behind human mobility. Thus, it is even harder to make prediction on the trajectory of extended time interval.

B. *Overview of Prediction Accuracy*

Now, we implement our designed method based on Markov model to make prediction on the trajectories. For the reason that our method make prediction by considering both temporal and spatial correlations, we employ the maximum predictability on the real entropy Π^{real} as a standard to evaluate the performance of our prediction. Since the trajectory we build consist of a user's 46 days location in the rate of one location per hour, we split the 1104 points trace into two parts. First part has 36 days' locations as train set, while the other part include 10 days' locations as test set.

For each user, we first train a Markov model based on train set to obtain the transition matrix. Then we utilize the



(a) Accuracy (b) Entropy S^{real} -Accuracy γ (c) Predictability Π^{real}-Accuracy γ

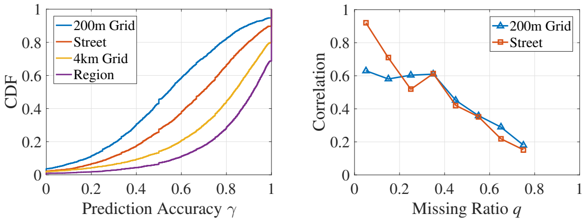
Fig. 6: **Performance of our prediction algorithm Correlated with entropy and prediction limits**

transition matrix to make prediction on the test set. Since our prediction of a certain time slot relies on its past n locations, we only make prediction when there are consecutive n known locations. Additionally, the following location should also be known if we desire to testify the prediction. Thus, we find the consecutive $(n+1)$ known locations during the test procedure and make prediction based on the former n data. By comparing the predicted result with the actual location, we can then evaluate our prediction algorithm. We denote the accuracy of prediction as γ:

$$\gamma = \frac{\text{number of accurate predictions}}{\text{number of total predictions}} \tag{14}$$

We first implement our method on the trajectory built on the scale of street. The distribution of prediction accuracy γ is shown in the Fig. 6(a). Over 20% of our predictions achieve the accuracy of 100%, which indicates that huge regularity implied in human movement can be fully explored, regardless of its appearing random. For the purpose of demonstrating whether our prediction method has approached the predictability, the relation between prediction accuracy γ and entropy S^{real} under the scale of street are manifested in a heat map in Fig. 6(b). A strong correlation between two variables can be observed. A trajectory with large entropy is supposed be difficult to predict since the entropy represents the uncertainty. Moreover, the relation between prediction accuracy γ and predictability Π^{real} is also illustrated in Fig. 6(c). An apparent linear relationship with most of data concentrated in the upper right corner is clearly shown, which exactly proves that our method makes full use of the regularity to achieve the predictability. The highest correlation coefficient of γ and Π^{real} reaches 0.82.

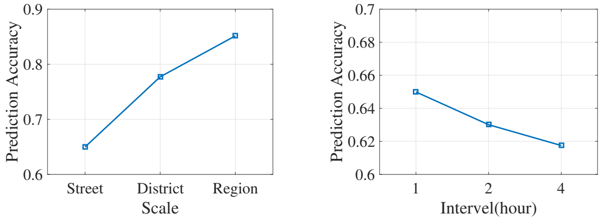
Then we analyze the effect of contextual information on prediction. The division of the street is regarded as context-based division since it is a combination of several sites which are close in location and similar in function, while the geographic grid provides no context of the location. Hence, we choose the two divisions of the same scale in contrary. Specifically, as for small scale, in comparison to the contextually clustered streets, we divide the city into 200m grids according to geographical coordinates. For the region scale, we choose 4km grid to be the contrary. The prediction accuracy of the four divisions are shown in Fig. 7(a). It is clearly illustrated that under both scales, the trajectories built on the context-based divisions have smaller entropy and larger predictability than those of geographic grids, which proves that the division with context can provide more information about the user's movement than division simply based on the geographical coordinates. In other words, context-based division preserves more information of a user being recorded in certain locations. Those preserved information, undoubtedly, will help us to explore the patterns of human mobility.



(a) Accuracy (b) Correlation Coefficient

Fig. 7: **Performance of prediction under different divisions**

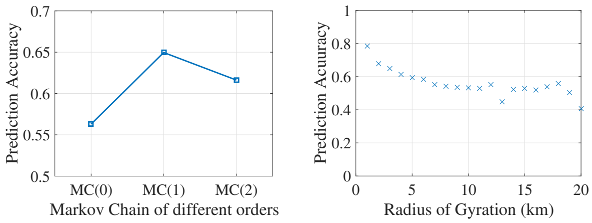
In order to study how correlation coefficient varies as the missing ratio q, we split the users into 10 groups according to their missing ratio in a interval of 0.1 and separately calculate the correlation coefficient of each group. The results under two divisions of the street scale are shown in the Fig. 7(b). As we can observe, the users with $q < 0.5$ have strongly positive correlations (correlation coefficient $r > 0.4$) between prediction accuracy and the predictability while the users with a considerable part of missing data have little correlations. The decrease of correlation may be explained by the reason that excessive missing data mask the actual uncertainty of the user's movement, resulting in the increase of the predictability that is practically hard to achieve. Also, in both small and large scales, the divisions with context have stronger correlations than the coordinate grid when the missing ratio $q < 0.3$, again indicating that prediction on context-based division has a better performance than on those without context.



(a) Spatial Resolution

(b) Temporal Resolution

Fig. 8: **Performance of prediction under different temporal and spatial resolution**



(a) Different Order

(b) Different Radius of Gyration

Fig. 9: **Influence of the order of Markov model and radius of gyration**

C. Analysis of accuracy under different situations

In this subsections, we analyze the prediction accuracy under various situations to explore the human mobility.

We first employ the prediction algorithm on the different spatial and temporal resolutions. For the spatial aspect, the prediction accuracy of three different resolutions are shown in the Fig. 8(a). The average of the prediction accuracy are listed in Table III. We can conclude from the results that prediction on the large scale is more accurate than prediction on the small scale, which is consistent with the result from entropy and predictability. As for the temporal aspect, we make predictions on the trajectories of 1 hour, 2 hours and 4 hours time interval. The average of prediction accuracy is illustrated in Fig. 8(b). The result of prediction accuracy shows the same characteristics as the entropy and maximum predictability: the extension in time interval leads to the decrease in accuracy.

Then, we analyze the influence of the order of Markov model and radius of gyration. We implement our prediction algorithm based on MC(0), MC(1) and MC(2) models. In other words, we make prediction of a user's movement based on different numbers of his past location data in trajectory. The accuracy of three models is presented in the Fig. 9(a). The prediction based on MC(1) achieve a higher accuracy than MC(2), which indicates that the human mobility actually holds little memories of the past location. A user may simply decides his next move based on his current location. As for the correlation between prediction accuracy and radius of gyration, we group the users according to their radius of gyration and calculate the average accuracy of each group. The result is shown in the Fig. 9(b). When the radius of gyration r_g is less than 10km, we can observe a decrease of accuracy as r_g increase. In comparison, when r_g reaches 10km, the accuracy remains steady regardless of increase in r_g, which indicates that the stationary users and non-stationary users may have different movement patterns.

TABLE III: Predictability and Accuracy under Different Division

Division	Accuracy	Predictability	Entropy
Street	67.07%	94.34%	0.58
200m Grid	54.63%	93.12%	0.67
Region	86.14%	96.80%	0.28
4km Grid	76.80%	95.20%	0.35

V. Related Work

In regard to human mobility, there have been some studies on calculating and estimating the predictability based on several stochastic models. Song *et al.* [1] derived the limit of predictability of mobility from entropy of a user's historical trajectory using Fano's inequality [10], [14]. Brockmann *et al.* [15] studied the travelling behaviors of human and established a mathematical model to describe the moving patterns based on Levy Flights, another probabilistic model proposed by Shlesinger [16], [17], [18], [19], [20]. Furthermore, some research link the human mobility with the social network, by segregating the similar users using the information from social media, more general and universal mobility patterns on a certain group of people can be extracted [2], [21], [22]. Also, research from Wesolowski *et al.* [23] suggested that mobility predictability are robust to the substantial biases in phone ownership across different geographical and socioeconomic groups. In summary, most of study have concluded that there exists huge regularity within a user's moving patterns and claimed that the human mobility is highly predictable. Our work shown in this paper, in addition to the previous studies, employs Markov models to make predictions, applying this predictability into practice.

There are also many researches have been conducted on the prediction of human mobility on various models, such as Markov Chain models [3], [4], neural network [5], Bayesian network [6], finite state machine [7]. Since the lack of standard test set on this problem, accuracy of a certain method is highly dependent on the location data used during the training and testing procedure. Lu et al. [3] implemented a Markov Chain model to analyze the travel patterns on mobile phone call data records by utilizing an O(1)-MC model and their accuracy of predictions achieved a average of 87% and 95% for the stationary trajectories and non-stationary trajectories, respectively. In comparison, another study conducted by Song *et al.* [24] on a campus-wide Wi-Fi wireless network indicated that the O(2)-MC model has the best performance with its accuracy around 65-72%.

It worth mentioning that most of the latest studies are based on evenly distributed and rarely missing ISP-collected dataset. As for researches on sparse trajectories, cluster-based methods and sampling-based methods are commonly used to alleviate the data sparsity. Zhang et al. [25] use a HMM-based group method to cluster users and predict on group level. Jeong et al. [26] propose a cluster-aided mobility predictor that leverage the similarities among users to improve prediction accuracy. Shokri et al. [27] use Markov model based on Gibbs sampling to model the human mobility, which is similar to our method. However, their model mainly focuses on trajectory matching, and they only use O(1)-MC model. Our work, on the other hand, aims at predicting users' movement based on partially missing app-collected dataset and restoring the missing data. Furthermore, we analyze the prediction accuracy under different situations, including different orders of the MC model, trying to uncover the implied moving patterns.

VI. Conclusion

Prediction based on users' historical trajectories are of great importance to improve the location based services. In this paper, aiming at addressing the three challenges of app-collected dataset, we carry out a comprehensive analysis of predictability and prediction method over a large scale app-collected location data. First, we employ a context-based segmentation method to preserve the physical context of app-collected data. Under the division, trajectories achieve the prediction accuracy 10% higher than those without context. Second, we design a Markov chain method based on Gibbs sampling in order to solve the unevenly distribution and the high missing rate. The prediction accuracy of the method reaches a high level with the correlation coefficient between prediction accuracy and predictability reaching 0.86. Third, we comprehensively analyze various factors that effect the prediction accuracy, including spatial and temporal resolution, orders of Markov models, radius of gyration. In general, Our analysis provides a systematic and comprehensive understanding of predictability of users' mobility patters on app-collected dataset.

References

[1] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[2] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. L. Barabasi, "Human mobility, social ties, and link prediction," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, Ca, Usa, August*, pp. 1100–1108, 2011.

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

[4] G. Liu and G. M. Jr, "A class of mobile motion prediction algorithms for wireless mobile computing and communications," *Mobile Networks and Applications*, vol. 1, no. 2, pp. 113–121, 1996.

[5] S. C. Liou and H. C. Lu, "Applied neural network for location prediction and resources reservation scheme in wireless networks," in *International Conference on Communication Technology Proceedings*, pp. 958–961 vol.2, 2003.

[6] S. Akoush and A. Sameh, "Movement prediction using bayesian learning for neural networks," in *International Conference on Systems and Networks Communications*, p. 6, 2007.

[7] J. Petzold, F. Bagci, W. Trumler, and T. Ungerer, "Global and local state context prediction," 2003.

[8] F. Xu, P. Zhang, and Y. Li, "Context-aware real-time population estimation for metropolis," in *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 1064–1075, 2016.

[9] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to english text," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1319–1327, 1998.

[10] RobertM, *Transmission of information : a statistical theory of communications*. M.I.T. Press, 1961.

[11] N. Navet and S. H. Chen, *On Predictability and Profitability: Would GP Induced Trading Rules be Sensitive to the Observed Entropy of Time Series?* Springer Berlin Heidelberg, 2008.

[12] C. P. Robert, G. Celeux, and J. Diebolt, "Bayesian estimation of hidden markov chains: a stochastic implementation," *Statistics [?] Probability Letters*, vol. 16, no. 1, pp. 77–83, 1993.

[13] M. K. Cowles and B. P. Carlin, "Markov chain monte carlo convergence diagnostics: A comparative review," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 883–904, 1996.

[14] A. K. Erlang, "The theory of probabilities and telephone conversations," 1909.

[15] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel. [nature. 2006] - pubmed result," 2006.

[16] M. F. Shlesinger, G. M. Zaslavsky, and U. Frisch, "Lvy flights and related topics in physics," *Lecture Notes in Physics*, vol. 450, 1995.

[17] M. C. Gonzlez, C. A. Hidalgo, and A. L. Barabsi, "Understanding individual human mobility patterns (nature (2008) 453, (779-782))," *Nature*, vol. 458, no. 7235, 2009.

[18] R. Cohen, S. Havlin, and D. Benavraham, "Efficient immunization strategies for computer networks and populations.," *Physical Review Letters*, vol. 91, no. 24, p. 247901, 2002.

[19] R. Metzler, A. V. Chechkin, V. Y. Gonchar, and J. Klafter, "Some fundamental aspects of levy flights," *Chaos Solitons Fractals*, vol. 34, no. 1, pp. 129–142, 2007.

[20] A. Rubio, V. Fras-Martnez, E. Fras-Martnez, and N. Oliver, "Human mobility in advanced and developing economies: A comparative analysis," 2010.

[21] M. C. GonzÁLez, C. A. Hidalgo, and A. L. BarabÁSi, "Understanding individual human mobility patterns," in *APS March Meeting*, pp. 779–82, 2008.

[22] N. Eagle, A. Clauset, and J. A. Quinn, "Location segmentation, inference and prediction for anticipatory computing.," pp. 20–25, 2011.

[23] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, and C. O. Buckee, "The impact of biases in mobile phone ownership on estimates of human mobility," *Journal of the Royal Society Interface*, vol. 10, no. 81, p. 20120986, 2013.

[24] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating next-cell predictors with extensive wi-fi mobility data," *IEEE Transactions on Mobile Computing*, vol. 5, no. 12, pp. 1633–1649, 2007.

[25] C. Zhang, K. Zhang, Q. Yuan, L. Zhang, T. Hanratty, and J. Han, "Gmove: Group-level mobility modeling using geo-tagged social media," in *KDD: proceedings. International Conference on Knowledge Discovery & Data Mining*, vol. 2016, p. 1305, NIH Public Access, 2016.

[26] J. Jeong, M. Leconte, and A. Proutiere, "Cluster-aided mobility predictions," in *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*, pp. 1–9, IEEE, 2016.

[27] R. Shokri, G. Theodorakopoulos, J. Y. L. Boudec, and J. P. Hubaux, "Quantifying location privacy," in *Security and Privacy*, pp. 247–262, 2011.