IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 10, 2020, accepted July 7, 2020, date of publication July 16, 2020, date of current version July 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009868

Measuring the Gap Between the Maximum Predictability and Prediction Accuracy of Human Mobility

JUNYAO GUO[1], **SIHAI ZHANG**[1,2], **(Senior Member, IEEE), JINKANG ZHU**[2], **(Life Member, IEEE), AND RUI NI**[3], **(Member, IEEE)**

[1]Key Laboratory of Wireless-Optical Communications, University of Science and Technology of China, Hefei 230022, China
[2]PCNSS, University of Science and Technology of China, Hefei 230017, China
[3]Wireless Technology Laboratory, 2012 Laboratory, Huawei Technologies Company Ltd., Shenzhen 518129, China

Corresponding author: Sihai Zhang (shzhang@ustc.edu.cn)

This work was supported in part by the Key Program of Natural Science Foundation of China under Grant 61631018, and in part by the Huawei Technology Innovative Research under Grant YBN2018095087.

ABSTRACT It has been claimed that human mobility is highly predictable and an upper bound of 93% predictability is achievable. However, there is a significant gap between the upper bound of predictability and the actual prediction accuracy in many data sets. This paper points out that this gap is caused by the difference between the user's actual distribution and the hypothesis in the derivation through the analysis on the upper bound of predictability. Then two statistics of the target user's mobility traces are proposed to measure this gap, whose effectiveness is validated by simulated traces and real-world data sets using five prevailing prediction models. The proposed MLP statistics can help with assessing the data quality and designing prediction algorithms. Our work makes the predictability upper bound become a more effective measure and extends the understanding of predictability research in human mobility prediction.

INDEX TERMS Human mobility predictability, real entropy, Call Detail Records.

I. INTRODUCTION

Understanding human mobility plays a critical role in various areas, such as urban planning [1], emergency management [2], public health [3], location-based services [4], personalized point of interest (POI) recommendations [5], economic forecasting [6], and transportation engineering [7]. Mobility research helps improve the user comfort and mobility prediction can benefit the network design and personal service optimization [8].

In the existing literature, various methods have been proposed to forecast human mobility, such as Markov chain models [9], neural networks [10], Bayesian networks [11], [12], data mining with different knowledge [13], [14] and effective time series prediction methods [15], [16]. Based on these models, there have been plenty of research works on mobility prediction, but the results and practical feasibility of the proposed predictive algorithms are somehow difficult to be generalized to the universal problem solving. In addition, people are not quite sure how well these algorithms will perform versus the best possible algorithm that could theoretically be constructed. So even for certain given data set, the best possible accuracy achievable and how well do the predictive algorithms perform is the very important topic.

The above question is answered partially, if not all, by Song *et al.* [17], who proposed mobility entropy to characterize the predictability of human mobility and pointed out that the probability of correctly predicting an individual's next location had an upper bound of 93% in their data set in 2010. Following this work, several papers have used this entropy metric to explore the limits of predictability in human dynamics [14], [18]–[25]. All these works have verified that the existing maximum predictability measures can reflect the potential prediction ability of different predicted objects. Specifically, the higher the maximum predictability of predicted objects, the higher the prediction accuracy.

Although Song *et al.* have proposed a pioneer model to calculate the maximum predictability of human mobility, knowing the upper bound has significant practical application

The associate editor coordinating the review of this manuscript and approving it for publication was Xiao-Sheng Si.

significance only when it is relatively tight, i.e. it is close to the extent which the actual optimal algorithm can reach. In fact, some literature reported that the existing maximum predictability measure may not provide accurate quantification of the real prediction ability for many data sets. Thus, this paper attempts to find when the maximum predictability is an accurate and valid measure, i.e. a tight bound.

The contributions of this paper are summarized as follows:

- We are among the firsts to study the gap between actual prediction accuracy and the theoretical bound of maximum predictability of human mobility. By using the CDR records with 192,805 users in a large Chinese city, we reveal the existence of the gap in real world traces considering a series of next-place prediction algorithms. We propose a theoretical formulation of this gap existence which is validated to be caused not by certain data set or the improper selection of prediction models.
- We propose two statistics, the probability of the most likely next position (MLP) and the standard deviation of location transition probability (SD), to measure the potential gap of users' maximum predictability. We quantitatively analyze the impact of the key statistics of the user's movement process on the gap. The MLP is validated to be an effective measure and beneficial to obtain the accurate boundary of the optimal predictor by the results of both simulated traces and real-world traces.

II. NOTATIONS AND DEFINITIONS

In this section, the concepts of mobility entropy and predictability in human mobility and how to calculate them are introduced.

A. NOTATIONS

For each target user, one location record is considered as a location access. After sampling and data cleaning, the location access records are sorted into time series of location access, i.e. trajectory. This sequence is modeled as a random process, and each location access is treated as a random state. The definitions and notations used in this paper are introduced below.

- $\boldsymbol{N}$: The size of state space, i.e. the number of distinct locations accessed by one user.
- $\boldsymbol{L} = \{x_1, x_2, \ldots, x_N\}$: The state pace, i.e. the location set with N distinct locations. $\boldsymbol{x_i}$ represents a possible location that the target user may access in L.
- $\boldsymbol{X_i}$: The ith state in random process, i.e. the ith location access of the target user.
- $\boldsymbol{h_n} = \{X_1, X_2, \ldots, X_n\}$: Historical trajectory arranged in chronological order with n location accesses.
- $\boldsymbol{T}$: A set of all mobility patterns that occur in one user's historical trajectory. $T_i \subset h_n$ represents a mobility pattern in T, i.e. a substring of historical trajectory.
- $\boldsymbol{T^j}$: The set of rank j mobility pattern, where j is the length of the pattern. $T_i^j = \{X_{i+1}, \ldots, X_{i+j}\}$ represents a member of T^j, i.e. a pattern with j location accesses.
- $\boldsymbol{MLP}$: A general concept that represents the probability that the user's next state is the most likely next position x_{ML}. $MLP(T_i)$ is the probability of x_{ML} corresponding to a specific mobility pattern T_i. $\overline{MLP}$ is a weighted average of $MLP(T_i)$ for all mobility patterns.
- $\boldsymbol{M_R}$: Random matrix whose transition probability is generated randomly and normalized to satisfy the constraints of the state transition matrix.
- $\boldsymbol{M_E}$: Equal probability matrix whose each transition probability equals $\frac{1}{N}$.
- $\boldsymbol{M_p}$: State transition matrixes with a set MLP of each transition state, $MLP \in [\frac{1}{N}, 1)$.
- $\boldsymbol{M_u}$: State transition matrixes of real users whose each transition probability is extracted from users' trajectory records.

B. DEFINITIONS OF MOBILITY PREDICTABILITY

In this section we explicitly introduce mobility predictions and predictability used in this work.

1) MOBILITY PREDICTION

Mobility prediction can be divided into individual mobility prediction and group mobility prediction. The individual mobility prediction studied in this paper refers to the prediction of the user's next position by extracting the movement characteristics from the individual's historical trajectory.

Based on the mobility prediction above, the definitions specified in [17] for predictability and its maximum theoretical value is introduced here.

2) PREDICTABILITY (Π)

The *mean probability* that an appropriate prediction algorithm can correctly predict a person's future whereabouts, given knowledge of all of the possible trajectories that could have led them to that point.

3) MAXIMUM PREDICTABILITY (Π^{max})

The highest potential predictability by assuming one possesses the best prediction algorithm that is theoretically possible, which is usually used to quantify the degree to which human activity is predictable.

III. RELATED WORKS

As to the research topics on prediction predictability in mobility prediction, there are mainly three questions: Is the current upper bound for one mobility prediction task achievable? What is the relationship between the upper bound of predictability and accuracy in actual prediction, and what are the factors affecting the two? Is the limit of predictability good enough for guiding the prediction accuracy?

A. RELATED WORKS ON PREDICTABILITY

Some works tried to design advanced prediction algorithms to approach the maximum predictability, but failed to obtain unified conclusion. Lu *et al.* [19] measure the maximum predictability of people in West Africa and find the theoretical limits are an approachable target. In [26], a modified Markov model is proposed for users who visit new locations in the test set, resulting in an improvement of 12.93% accuracy. In [13], the Diffusion kernel model based on propagation network obtained better performance than current models. But the prediction performance in [26] and [13] can not approach the upper bound of predictability of used data sets.

Besides, an important topic is to explore the relationship between the limits of predictability and prediction accuracy. Reference [27] models the relationship between accuracy and predictability for specific algorithms using students mobility data by composite Gaussian function, but this conclusion needs to be confirmed by more data sets. Reference [13] reported that users with the same predictability correspond to different prediction accuracies, which might be caused by the limits of predictability being not a tight bound. Another interesting topic is to find out the critical factors affecting the maximum predictability and prediction accuracies, such as the temporal and spatial resolution of data [28], [29], the preference of exploration [25] and the data quality. The differences of mobility predictability of college students between gender, age and grade groups using campus consumption records is analysed in [28]. The upper limit of predictability increases as the temporal resolution becomes finer-grained in the sensor records of 14 participants [29]. The extensive prediction performance span is attributed to human exploration of new locations [25].

Some other researchers have tried different methods or modified predictability measures to quantify the predictability of human mobility, such as mutual information [30], instantaneous entropy [31], and so on. Smith *et al.* [32] achieve a tighter upper bound using the Geolife data by considering real-world topological constraints. But this method required more detailed spatiotemporal data, continuous GPS information, and real world geographical location. So the redefined upper bound is more difficult to popularize than the definition in [17]. In addition, [23] pointed out that the approximation algorithm (Lempel-Ziv estimator) of the original predictability in [17] may lead to the overestimation of predictability, and gave the inherent deviation of four different sequences.

B. BASELINE PREDICTION MODELS

Four baseline prediction models are introduced here which are chosen for the performance comparison in this work. These four models have been tested in existing works [13], [16] and they have different characteristics. For example, DK model is similar to the neural network prediction model and gradient descent operations are involved in the calculation without requiring large amounts of training data. ST model takes into account the factors of moving time intervals, and performs better than models that do not consider time characteristics.

The prediction accuracy to assess the performance for one specific algorithm is defined as the number of correct predictions n_{correct} over the total number of predictions n_{total}:

$$Acc^{alg} = \frac{n_{\text{correct}}}{n_{\text{total}}} \tag{1}$$

And the gap between the accuracy of a specific algorithm and the maximum predictability can be calculated by (2).

$$Gap^{alg} = \Pi^{max} - Acc^{alg} \tag{2}$$

1) MOST FREQUENT VISIT MODEL (MFV)

In MFV model [14], the probability of next check-in at location x of MFV is defined as (3).

$$P^{MFV}(X_{i+1} = x|X_i = x_i, h_i = h) = \frac{|\{X_r|X_r \in L, X_r = x\}|}{|\{X_r|X_r \in L\}|} \tag{3}$$

2) FALLBACK MARKOV CHAINS (FMC)

The n-order Markov chain means that an individual's next place is related to the previous n visited locations. The ''fallback'' Markov chain uses the results of the n-order Markov model when it makes a prediction, or results of the MFV predictor if the Markov predictor has no prediction. It is reported that the n-order Markov models have higher complexity and lower accuracy when $n > 2$ [9], [19], so 1-order and 2-order FMC are chosen to implement next place prediction in this work.

3) DIFFUSION KERNEL MODEL (DK)

Diffusion kernel model was originally used to describe the internal thermal diffusion of manifold nuclei [33] and was later applied to the prediction field [13]. The mobility behavior is mapped to the heat diffusion process in the hidden space, and the prediction is based on the magnitude relationship of the Euclidean distance between the nodes.

4) SPATIAL TEMPORAL MODEL (ST)

The Spatial Temporal model is a slot-based continuous next-place prediction model [16]. The model first determines whether the next time slot is a resident state or a moving state based on the residence time of various spatial and temporal characteristics, and then makes a position prediction.

IV. METHODOLOGY

In this section, the concepts of mobility entropy and predictability in human mobility and how to calculate them are introduced.

A. DERIVATION OF PREDICTABILITY

Before analysing the gap between the bound of predictability $\Pi^{\max}$ and actual prediction accuracy, the derivation of predictability Π done in [17] should be reviewed first, as below.

$P(X_n|h_{n-1})$ denotes the conditional probability distribution for the target user's nth location state given h_{n-1}. The best prediction from any predictor is to return the most likely next state x_{ML} from the distribution. In (4), $\pi(h_{n-1})$ denotes the probability of x_{ML} and $P(X_n = x|h_{n-1}) = P(x|h_{n-1})$ is the probability that the user will select x as the next state X_n given the history h_{n-1}.

$$\pi(h_{n-1}) = \sup_x P(X_n = x|h_{n-1}) = P(x_{ML}|h_{n-1}) \tag{4}$$

Let $P_a\left(\hat{X}_n|h_{n-1}\right)$ be the distribution generated by an arbitrary prediction algorithm α over the next possible state $\hat{X}_n$. So the probability of correctly predicting the user's next state is

$$Pr_a\left(X_n = \hat{X}_n|h_{n-1}\right) = \sum_x P\left(x|h_{n-1}\right) P_a\left(x|h_{n-1}\right). \tag{5}$$

There is $\pi\ (h_{n-1}) \geq P\left(x|h_{n-1}\right)$ for any x, so

$$\begin{aligned} Pr_a &\left\{X_n = \hat{X}_n|h_{n-1}\right\} \\ &= \sum_x P\left(x|h_{n-1}\right) P_a\left(x|h_{n-1}\right) \\ &\leq \sum_x \pi\ (h_{n-1})\, P_a\left(x|h_{n-1}\right) = \pi\ (h_{n-1}). \end{aligned} \tag{6}$$

Since $\Pi(n)$, defined in (7), is the highest prediction success rate of user at nth location state, the overall predictability Π can be regarded as the average predictability of time series.

$$\Pi(n) = \sum_{h_{n-1}} P\left(h_{n-1}\right) \pi\ \left(h_{n-1}\right). \tag{7}$$

$$\Pi \equiv \lim_{n\to\infty} \frac{1}{n} \sum_i^n \Pi(i). \tag{8}$$

B. MEASUREMENTS OF MAXIMUM PREDICTABILITY

To compute the upper bound of predictability $\Pi^{\max}$, [17] relate user's entropy to Fano's inequality. The entropy of the time series can be expressed as (9).

$$S = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} [\sum_{h_{i-1}} P(h_{i-1}) S(X_i|h_{i-1})] \tag{9}$$

The conditional entropy $S(X_n|h_{n-1})$ of target user over nth state given h_{n-1} is written as (10), where $p = \pi(h_{n-1})$.

$$\begin{aligned} S(X_n|h_{n-1}) &= -\sum_x P(x|h_{n-1})\log_2 P(x|h_{n-1}) \\ &= -p\log_2(p) - \sum_{x\neq x_{ML}} P(x|h_{n-1})\log_2 P(x|h_{n-1}) \end{aligned} \tag{10}$$

Since the actual next state distribution $P(X_n|h_{n-1})$ is unknown in practice, they define $S(X'_n|h_{n-1})$, an upper bound for $S(X_n|h_{n-1})$, with a new distribution $P(X'_n|h_{n-1})$.

$$P(X'_n|h_{n-1}) = (p, \frac{1-p}{N-1}, \frac{1-p}{N-1}, \ldots, \frac{1-p}{N-1}). \tag{11}$$

$$\begin{aligned} S(X'_n|h_{n-1}) &= -plog_2(p) - (1-p)log_2(\frac{1-p}{N-1}) \\ &= -plog_2(p) - (1-p)log_2(1-p) \\ &\quad + (1-p)log_2(N-1) \\ &\equiv S_F(p) = S_F(\pi(h_{n-1})). \end{aligned} \tag{12}$$

$S(X_n|h_{n-1}) \leq S(X'_n|h_{n-1}) = S_F(\pi(h_{n-1}))$ represents an appropriate rewriting of Fano's inequality [34]. And then assuming $\Pi = \Pi^{\max}$ is the condition of $S(X_n|h_{n-1}) = S_F(\pi(h_{n-1}))$, one can find the maximum predictability $\Pi^{\max}$ satisfies (13).

$$S = S_F\left(\Pi^{\max}\right) \tag{13}$$

Finally, the calculation of $\Pi^{\max}$ can be obtained by entropy S as (14) and (15). $\Pi^{\max}$ gives the upper limit of the accuracy of any prediction algorithm.

$$S = H(\Pi^{\max}) + (1-\Pi^{\max})log_2(N-1). \tag{14}$$

$$H(\Pi^{\max}) = -\Pi^{\max}log_2(\Pi^{\max}) - (1-\Pi^{\max})log_2(1-\Pi^{\max}). \tag{15}$$

C. THREE ENTROPY

From the spatiotemporal correlation inside the mobility trajectory, three entropy measures are proposed in [17].

Random Entropy: $S^{\text{rand}} \equiv log_2 N$.

Temporal-Uncorrelated Entropy: $S^{\text{unc}} \equiv -\sum_{i=1}^{N} p(x_i)$ $log_2\, p(x_i)$, where $p(x_i)$ is the probability that x_i occurs in the history trajectory.

Real Entropy: the real entropy S^{real} is defined by $-\sum_{T_{i'}\subset h_i}$ $p(T_{i'})log_2 p(T_{i'})$, where $p(T_{i'})$ is the probability of finding a particular mobility pattern $T_{i'}$ in the trajectory h_i.

The *real entropy* not only depends on the access frequency, but also considers the order of location access and the duration at each location. Since among the three entropy, the real entropy is closest to the uncertainty of the user's mobility, Π^{real} is coined as $\Pi^{\max}$. The Lempel-Ziv data compression [35] is used to calculate the S^{real}. For a time series with n states, the entropy is estimated by (16), where Λ_i is the length of the shortest substring at state i which does not previously appear from state 1 to $i-1$.

$$S^{est} = \left(\frac{1}{n}\sum_i \Lambda_i\right)^{-1} \ln n \tag{16}$$

V. TWO KEY STATISTICS TO ANALYSIS GAP

In this section, the derivation of the maximum predictability is discussed to reveal the inherent reason for the gap existence. Then two key statistics of mobility process are proposed to explore the factors that affect this gap.

Although [17] obtained the relationship between the upper bound of predictability and the entropy $\Pi \leq \Pi^{\max}(S, N)$ through Fano's inequality and Jensen's inequality, however, it is obvious that not all of the actual distributions of users to select the next state is the same as (11). Thus when applied to practical predictions, this upper bound of predictability may not be a perfect indicator for the optimal prediction accuracy.

As shown in Fig. 1, the assumed distribution in [17] results in an entropy greater than or equal to that of distribution of the real world traces, which results in an overestimated upper bound on predictability. Reference [32] eliminated the unreachable locations using the real-world constraints, and applied uniform distribution to fewer possible locations. This

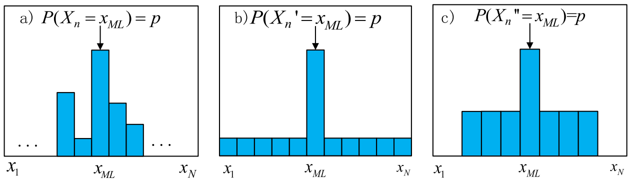


FIGURE 1. (a) The actual distribution (unknown in practice) over next state. (b) Distribution used in [17]. (c) Refined distribution in [32].

approximation has an entropy lower than [17] but higher than the actual entropy. However, this method requires a large amount of external information except trajectory.

Then, according to the original derived assumptions, two statistics of the user's historical location access distribution are presented to measure the actual impact of the overestimation. Given a particular history movement pattern T_i, the transition probability of x_{ML} in (8) is coined here as $MLP(T_i)$, and the Standard Deviation of transition probability on the remaining $N-1$ positions in (17) is coined as $SD(T_i)$.

$$SD(T_i) = \sqrt{\frac{\sum_j (Pr(x_j|T_i) - Mean^2[Pr(\hat{x} \neq x_{ML}|T_i)]}{|\{\hat{x} \neq x_{ML}\}|}} \quad (17)$$

$$Mean[Pr(\hat{x}|T_i, \hat{x} \neq x_{ML})] = \frac{\sum_j Pr(x_j|T_i)}{|\{\hat{x} \neq x_{ML}\}|} \quad (18)$$

Since the user's movement process contains a lot of patterns, the weighted average of $MLP(T_i)$ and $SD(T_i)$ of all history movement patterns coined here as $\overline{MLP}$, $\overline{SD}$ respectively. $\overline{MLP}$ can measure the user's overall prediction decision, as shown in (19).

$$\overline{MLP} = \sum_{T_i \subset h_{n-1}} P(T_i)\pi(h_i) = \sum_{T_i \subset h_{n-1}} P(T_i) MLP(T_i) \quad (19)$$

Similarly, $\overline{SD}$ can be calculated by (20).

$$\overline{SD} = \sum_{T_i \subset h_{n-1}} Pr(T_i) SD(T_i) \quad (20)$$

According to the rank of mobility patterns, (19) can be written as (21).

$$\overline{MLP} = \sum_{j=1}^{n-1} Pr(T^j)[\sum_{T_i^j \subset h_{n-1}} Pr(T_i^j) MLP(T_i^j)] \quad (21)$$

According to the definition above, two understandings are pointed here. First, the bigger the $MLP(T_i)$, the smaller the probability of the remaining $N-1$ positions. When the $MLP(T_i)$ is large enough, the difference between the assumed distribution and the original distribution is indistinguishable, and the degree of overestimation is significantly reduced. Second, when $SD(T_i)$ approaches 0, the probability of overestimation is negligible. Therefore, these two statistics are expected to show that what kind of data set has an attainable predictability upper bound, that is, under what circumstances, the gap generated by overestimation of the predictability is slightly negligible.

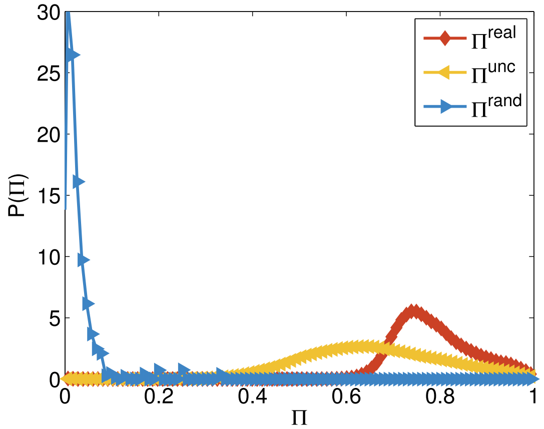


FIGURE 2. The distributions of the upper bound of predictability calculated by three entropies.

VI. EXPERIMENTS OF MOBILITY TRACES

This section describes the two types of data sets used in this paper, based on which the gap phenomenon is presented. Wireless technologies allow us to sense and collect massive repositories of spatio-temporal data [36] and most of the existing works are based on the data collected from wireless communication systems, such as Call Detail Records (CDR), WiFi (Wireless Fidelity) logs, GPS (Global Positioning System) traces, social media data [37] etc. The CDRs are chosen in this paper for the following reasons. First, as a by-product of routine billing by telecommunication service operators, CDR data can provide users' location information at a rather lower cost on a larger scale, which is appropriate to represent human mobility. Second, the original predictability theory of human mobility is performed using the CDRs in [17].

A. REAL WORLD TRACES

The CDR data set is provided by one telecom operator in China, which is collected for 6 months, 184 days, from July 1, 2014, to December 31, 2014. 194,336 anonymous mobile phone subscribers registered in one city are included in this data set. The user's location is provided as the location area code (LAC). There are 270,932,374 records from 453,752 locations in the data set. Each record has 12 features, and four of which, SERVICE NBR, START TIME, ROAM CITY ID, LAC ID, are extracted for predictability calculation and next place prediction.

The number of location updates of most users is distributed between 500 and 2,000. Users with fewer than 500 updates are filtered out because their records are not enough to reflect the behavior characteristics of the observation period. The number of unique locations visited, N, represents the visiting patterns of users. 80% of users' visiting limits in 100 locations, and minority of users visit more than 150 locations.

In this paper, track points are resampled to eliminate the ping-pong effect. Specifically, the trajectory point (x_i, t_i) is said valid if $x_i \neq x_{i-1}$ and $t_i - t_{i-1} \geq \triangle t$, $\triangle t = 30\ min$, to eliminate invalid records generated by users frequently switching between adjacent base stations.

B. SIMULATED TRACES

Human mobility has been claimed to follow *Levy walk* characteristics [38], which tells that the movement of target user is of a short distance around a location at most of the time, but in a few cases there is also a long distance movement. In practical cases, Markov related models are often used to model the movement process, such as Semi-Markov model, Finite Markov chains [39], Gaussian Markov model [40] etc. In this paper, the Markov state transition matrix defined in the state space is adopted to represent the visiting behavior or visiting preference of users and generated random visiting sequences are used for verification. Without loss of generality, only the 1-order state transition matrix is taken into consideration.

Among the four kinds of state transition matrixes used in this paper, M_R, M_E and M_p are generated randomly according to the set rules, while M_u is calculated from the historical trajectory of each real user. Each state transition matrix represents an access distribution for a target user. When generating random access sequences, the target user's movement in space is regarded as the random movement between possible locations according to the state transition matrixes. M_R and M_E are designed to observe the generation of gap at the boundary condition of theoretical derivation ($MLP = \frac{1}{N}$). M_p is used to observe the change trend of gap of sequences with different *MLP* eigenvalues. M_u is used to compare the performance of real users' simulation models to real world trajectories.

The parameter settings of the state transition matrixes are shown in Table 1, which helps to understand the generation of visiting sequences with different characteristics of *MLP* and *SD*. The generation algorithm of M_p is shown in Algorithm 1. M_{p1}, M_{p2} and M_{p3} are simulation matrixes of M_p with different parameters N = 50, Len = 50,000; N = 10, Len = 50,000 and N = 50, Len = 2,000. In particular, M_u denotes the real state transition matrixes of all users from CDR data set, where the N, *MLP* and *SD* of each M_u are calculated from real world CDR traces.

Note that the validation on the simulated traces confirms that the N of the generated trajectory under the parameter setting of Table 1 is consistent with the state space size set. In addtion, for the traces generated by M_u, the distribution of predictability and prediction accuracy are consistent with the distribution of the real trajectory of the user set. These validation is not demonstrated in this paper due to page limitations.

C. EXPERIMENT VERIFICATION OF GAP

The gap between actual prediction accuracy and theoretical limit of predictability of two kinds of mobility traces is discussed in this subsection.

1) EMPIRICAL GAP IN REAL WORLD TRACES

The analysis of the upper bound of predictability reveals that the theoretical upper bound of average prediction accuracy of the user set in CDRs can be as high as 80%. Five mainstream forecasting models, 1-order FMC, 2-order FMC, DK, MFV and ST, are used to predict the next place of real world users. The track set of each user is divided into training set and test set at a ratio of 5:1, which are used to perform model training and next position prediction separately for each target user.

Algorithm 1 M_p Generation Algorithm

```
Input:
N : Matrix dimension
p: MLP of each transition state
Output: $M_p$
Initializes an $N \times N$ dimensional matrix $M$
for $i = 0$ to $N - 1$ do
    Select a random integer $s$ from the range $[1, N)$
    for $j = 0$ to $N - 1$ do
        if $j == s$ then
            $M(i)(j) \leftarrow p$
        else
            $M(i)(j) \leftarrow \frac{1-p}{N-1}$
        end if
    end for
end for
return $M_p \leftarrow M$
```

The prediction results of all models and the distribution of users' predictability are shown in Fig. 3(a), which tells the following findings. Firstly, the predictive performance of all models varies no more than 10% while FMC(1) achieves the highest prediction accuracy at 0.62. Secondly, in this user set, the span of users' prediction accuracies is large, unlike $\Pi^{\max}$, which is concentrated distributed around a high value. so there is a big gap between actual prediction result and maximum predictability in the user set. The gap between the accuracy and predictability can be calculated by (2) and shown in Fig. 3 (b) that the FMC(1) has the smallest gap, while the MFV model has the largest gap.

From the empirical results above, it is proved that the existence of gap in real world data set is a non-accidental phenomenon and is not caused by the poor choice of prediction algorithms.

2) GAP IN SIMULATED TRACES

It has been described in Section IV that $\Pi^{\max}$ is the average probability that the theoretical best predictor makes correct predictions while the best prediction from any predictor is to return the location with the highest probability from the real distribution of next location. The real distribution of location access of simulation objects is 1-order Markov chain, so the 1-order Markov Model is used to predict the generated location visiting sequence.

TABLE 1. Parameters and statistics of transition matrixes M_R, M_E, M_p, M_u.

M	N	Len	MLP	SD	Comment
M_R	50	50,000	≈1/N	≈0	Generate Simulated Traces
M_E	50	50,000	1/N	0	Generate Simulated Traces
M_p	10; 50	2000; 50,000	0.02 : 0.1 : 0.92	0	Generate Simulated Traces
M_u	(0, 100]	2,000	[0.2332, 1]	[0, 0.2232]	Extracted from Real Traces

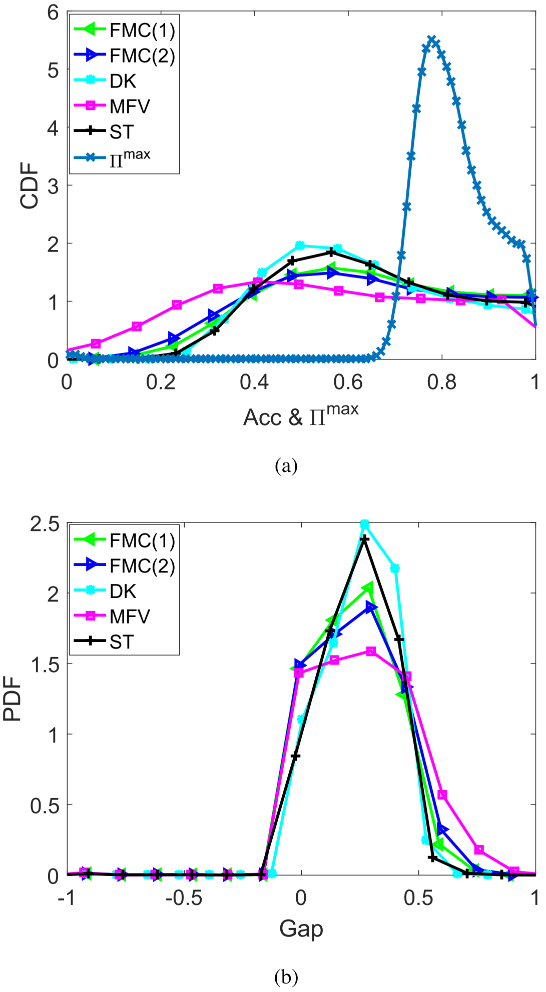


FIGURE 3. (a) Probability distribution of the maximum predictability and the prediction accuracy of the user set. Mean values are 0.6262, 0.6120, 0.6054, 0.5514, 0.6232, 0.8241. (b) Probability distribution of the gap of all models.

According to (11), the boundary condition $MLP = \frac{1}{N}$ is considered a special case to explore firstly. The result of M_R is similar to M_E, which turns out that the probability distribution of M_R is approximately uniform. The following analysis only takes M_E as an example. The average prediction accuracy of M_E as shown in Fig. 4 is only 0.0254, which is approximately the probability of random selection of $N = 50$ states. Calculated by (14) and (15), the maximum predictability Π^{max} is in [0.5015, 0.5590] while Π^{rand} is 0.0212. The prediction result is far from Π^{max}, so there is a gap between the maximum predictability and prediction accuracy of M_E.

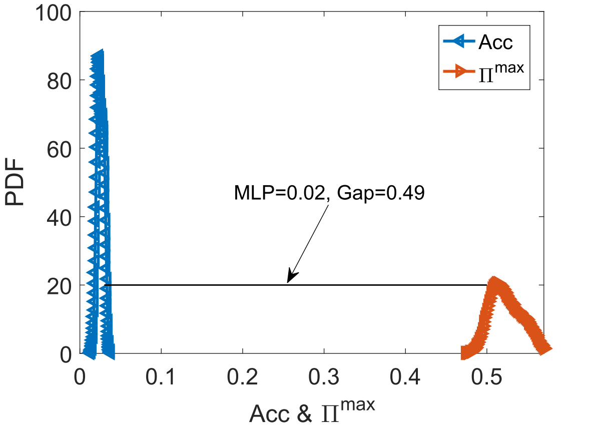


FIGURE 4. The gap between the maximum predictability and prediction accuracy of M_E.

VII. MEASURING THE GAP BY KEY STATISTICS

In this section, we verify that there is a linear relationship between the *MLP* and the gap of mobility traces.

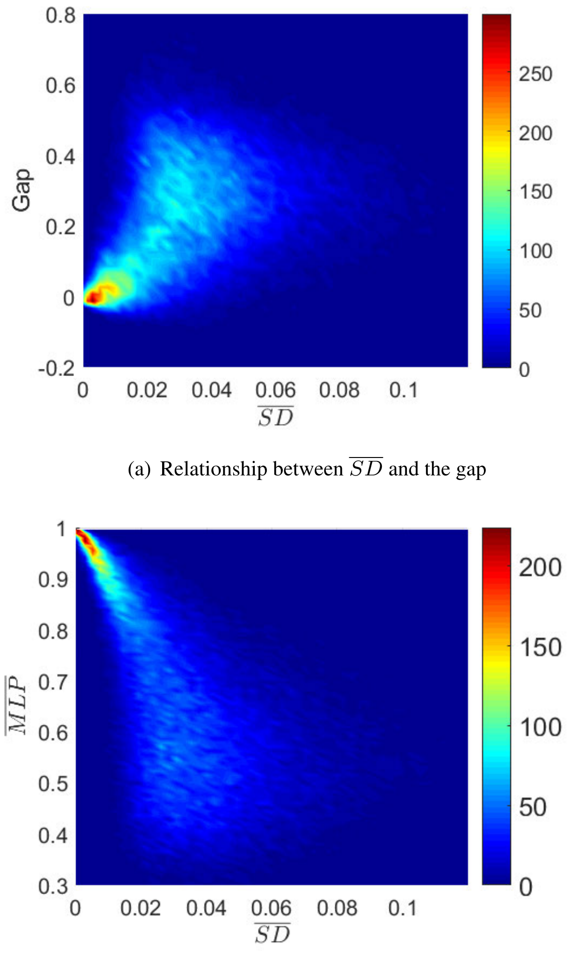
A. RELATIONSHIP BETWEEN STATISTICS AND GAP

1) KEY STATISTICS OF REAL WORLD TRACES

For each user in CDRs, $\overline{MLP}$ can be calculated from the historical trajectories by (19). After mapping each user's $\overline{MLP}$ to its predictability upper bound and prediction accuracy, We can analyze the relationship between gap and $\overline{MLP}$, as shown in Fig. 6 (b). The larger the $\overline{MLP}$ of a user's location visiting distribution, the smaller the gap between the maximum predictability and prediction accuracy. And the relationship between the two is almost a linear decreasing relationship.

Similarly, the results of $\overline{SD}$ calculated by (20) are shown in Fig. 5. Firstly, Fig. 5 (a) shows that when $\overline{SD}$ is close to 0, the gap is also close to 0, i.e. the overestimation of predictability upper bound is almost nonexistent, which is consistent with theoretical analysis. With the increase of $\overline{SD}$, the gap shows an overall increasing trend, but the gaps with the same level of $\overline{SD}$ are not the same when the span is large. Secondly, Fig. 5 (b) shows that when $\overline{MLP}$ is close to 1, the $\overline{SD}$ is close to 0. When $\overline{MLP}$ decrease from 1, the corresponding $\overline{SD}$ values and the range of which keep increasing.

Both $\overline{MLP}$ and $\overline{SD}$ are calculated from the probability distribution of the next location access. However, it can be seen from the above results that $\overline{MLP}$ is more suitable as an indicator of gap, and it has a greater impact on gap. Because the distribution of the same $\overline{SD}$ may be different, while $\overline{MLP}$ is the key factor affecting user's location access and the



(a) Relationship between $\overline{SD}$ and the gap

(b) Relationship between the $\overline{SD}$ and $\overline{MLP}$

FIGURE 5. The impact of $\overline{SD}$ on the gap between the average accuracy and theoretical maximum predictability of real world traces of 190K users.

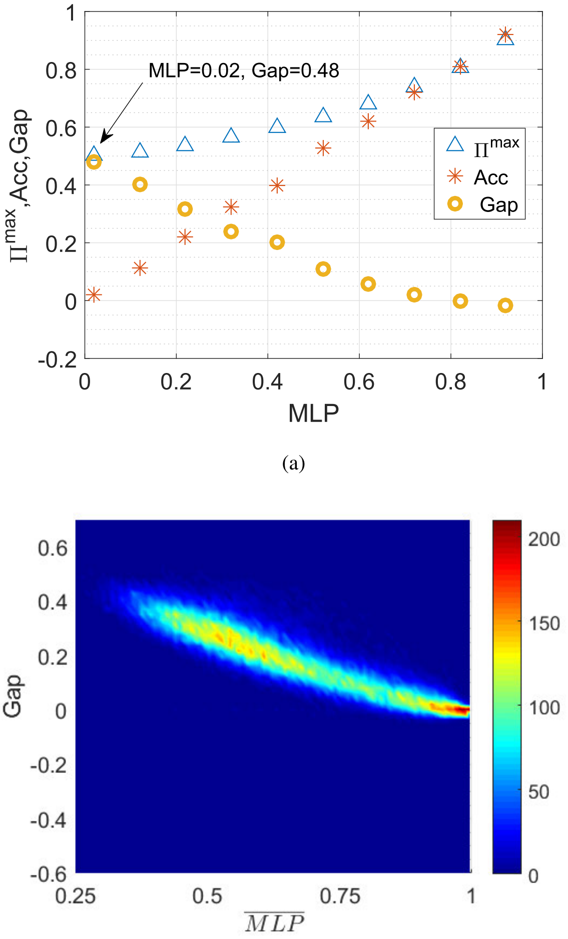
prediction decision. Since $\overline{SD}$ is not suitable for measuring and distinguishing the gaps of users, subsequent discussions will no longer be conducted around $\overline{SD}$.

2) KEY STATISTICS OF SIMULATED TRACES

The sequences with different *MLP* can be generated by M_p, and the corresponding gaps calculated by (2) are shown in Fig. 6 (a). We conclude that for $MLP \in [\frac{1}{N}, 1)$, the smaller the *MLP* is, the bigger the gap between actual prediction accuracy and the limit of predictability is. And when $MLP = \frac{1}{N} = 0.02$, the gap of M_{p1} is largest at 0.48, which is basically consistent with the results of M_E.

Since the generated simulation model is difficult to fully simulate the user's actual location access, this paper also simulates random visitation sequence according to the estimated location access distribution of real users, M_u. The results of M_u will be analyzed and compared with the results of users' real trajectories in subsection VII-B.

In summary, by analyzing the relationship between key statistics and gap, the key issues without consensus, whether the upper bound of predictability is reachable, can be answered. That is, the data set with $MLP \rightarrow 1$ and $SD \rightarrow 1$ has an attainable predictability upper bound and as *MLP* approaches the boundary of $\frac{1}{N}$ from 1, predictability upper bound appears increasingly unreachable.



(a)

(b)

FIGURE 6. (a) The relationship between *MLP* and the gap of M_{p1}. (b) Relationship between $\overline{MLP}$ and the gap between Π^{max} and accuracy achieved by FMC(1) of 190K users.

B. IMPACT OF MLP ON GAP

In this part, we quantitatively analyze the impact of $\overline{MLP}$ on the gap by comparing the linear relationship between $\overline{MLP}$ and gap in the results of simulated traces and real world traces.

The $\overline{MLP}$ of the real user set $\in$ [0.2332, 1] are divided into 16 intervals with the interval length = 0.05. The average of the gaps in the same $\overline{MLP}$ interval is seen as the gap corresponding to the interval's upper boundary. For example, the average of all gap values in (0.20 0.25] is recorded as the gap at $\overline{MLP} = 0.25$.

Fig. 7 presents the linear relationships between $\overline{MLP}$ and the gaps in two kinds of data sets, which tells the following conclusions. First and most importantly, $\overline{MLP}$ is verified to be a good indicator for calculating or estimating the gap in an approximately linear relationship. Second, when dealing

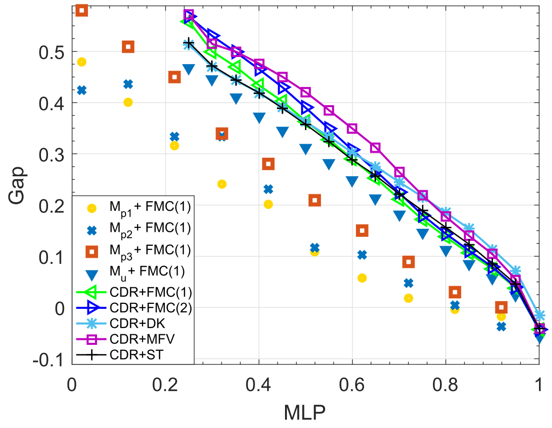


FIGURE 7. The linear relationship between $\overline{MLP}$ and the gaps achieved by different data sets and prediction models. M_p1, M_{p2} and M_{p3} are simulation matrixes with different parameters N = 50, Len = 50,000; N = 10, Len = 50,000 and N = 50, Len = 2,000.

with real data set, the gaps of these five prediction models follow a similar distribution. In addition, when using the parameters calculated from the data of actual users, the results of simulation model M_u are extremely consistent with that of other prediction models. Finally, the results of real data set achieved by all prediction models have obvious larger gap compared with the theoretical simulation model M_p, which is probably because the sequences generated by fixed parameters are difficult to simulate the visiting sequences of the whole user set.

The random visitation sequence generated by simulation is affected by three factors: distribution characteristics presented by MLP, state space size N, and sequence length Len. From the results of M_{p1}, M_{p2} and M_{p3}, we can find that N has less influence on the linear relationship while Len has more significant influence. When L of M_p is close to the trajectroy length of the real world user set, the result of simulated M_{p3} is closer to the result of M_u. The slight deviation between the two can be explained by two factors. On the one hand, the number of visiting locations and the trajectory length of each user in the user set are different. On the other hand, the location visitation in the test set may be different from the training set.

In conclusion, the indicator MLP can quantify the gap between the performance of theoretical optimal predictor of the dataset and the upper limit of predictability.

C. INFLUENCE OF RANK OF KEY STATICS

The definition of $\overline{MLP}$ is further discussed in this section. The currently calculated $\overline{MLP}$ only mines the predictive power of the target user's rank 1 mobility patterns. Further, the influence of the mobility patterns of different ranks on the calculation of the $\overline{MLP}$ and the relationship between the $\overline{MLP}$ and the gap can be explored.

$$\overline{MLP} = \sum_{j=1}^{k} Pr(T^j)[\sum_{T_i^j \subset h_{n-1}} Pr(T_i^j)MLP(T_i^j)] \qquad (22)$$

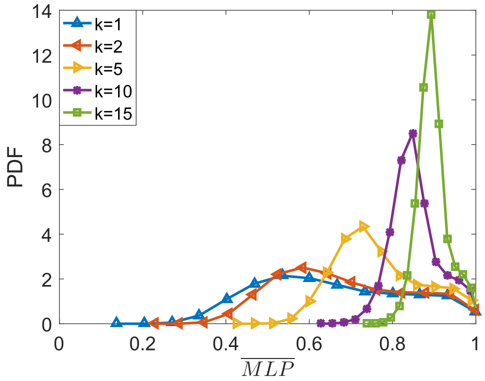


FIGURE 8. Probability distribution of $\overline{MLP}$ in the user set corresponding to different ranks k.

Corresponding to (22), the higher the rank k, the larger the range of j, $j \in (1, k)$ and $k \leq n-1$. The larger j is, the smaller $Pr(T^j)$ is and more high-order movement patterns are mined. But $Pr(T^j)$ decreases only at a constant rate, and the corresponding $MLP(T_i^j)$ will rise rapidly to approach 1, so the value of $\overline{MLP}$ increases as the rank k increases. When most $MLP(T_i^j)$ are close to 1, the $\overline{MLP}$ will also approach 1.

The cases of k = {1, 2, 5, 10, 15} are explored as Fig. 8. Same as the previous analysis, the larger the k, the larger the $\overline{MLP}$. In experiment, when $k \geq 20$, the $\overline{MLP}$ of most users are close to 1.

It should be noted that although a higher rank leads to a larger $\overline{MLP}$, it does not mean that the $\overline{MLP}$ is more effective in measuring the gap. As the rank increases, the relationship between $\overline{MLP}$ and gap no longer satisfies the linear relationship. Although with a higher rank, more mobility patterns and correlations are mined, the distribution difference between the training set and the test set becomes larger which results in lower robustness. In practical applications, the appropriate rank should be chosen to calculate the $\overline{MLP}$.

VIII. CONCLUSIONS

This paper starts from the observation that there is a very significant gap between the upper bound of predictability and the actual prediction accuracy. This motivates us to study and models the prediction accuracy, and the maximum predictability of human mobility, then to derive the gap between such prediction accuracy and such maximum predictability of human mobility. We revisit the theory behind the maximum predictability and derive two statistics to measure such gap. The gap and the validity of the metric MLP are verified by the results of simulated traces and real world trajectories. Simulated traces can verify the impact of MLP on the theoretical gap. Because both the prediction model and the trajectory generation model are Markov Models, the prediction decision is almost returning results from the actual distribution of location visits, which achieves the performance of theoretically optimal predictor. Besides, the real world trajectories are extracted from a large CDR dataset of 6 months in a city of China, with about 200,000 users and more than

450,000 locations, which can reflect the gap in the mobility prediction of the actual situation.

In particular, section III details that there is currently no consensus on whether the upper bound of predictability is reachable. Firstly, this paper answers this key question from the perspective of data set characteristics. That is, the data set with $MLP \rightarrow 1$ has an attainable predictability upper bound. Then, in view of the case that the predictability upper bound of the data set is not reachable, i.e. the overestimation, this paper illustrates the possible range of gap. Finally, the impact of *MLP*, a general and interpretable indicator, is discussed to quantify the gap between the theoretical optimal predictor of the data and the upper bound of predictability.

Compared with the current work, the method proposed in this paper proposes key statistics to quantify the gap between the existing theoretical upper bound and the achievable limit to obtain a more accurate and effective theoretical limit. We point out the problem of traditional research that there is still no unified understanding of whether the upper bound of predictability proposed by Song et al [17] is attainable. Literature [32] proposed a tighter upper limit, but requires precise, continuous spatio-temporal latitude and longitude information which is often difficult to obtain. In addition, the accuracy of the modified upper bound depends on the granularity of the spatiotemporal information. While the proposed method in this paper only requires the location records of the target object (whether it is area label or latitude and longitude information), our work is more conducive to popularity and application.

This work enables the prediction design to have a credible target $\Pi^{\max} - gap$, which can be applied to assess data quality during data collection and data set selection and establish the prediction objective. It can also help to select and design proper prediction algorithms. For example, if the performance of the one prediction algorithm is far from $\Pi^{\max} - gap$, another more suitable model should be considered because the possible bound is not yet reached. Otherwise, such poor prediction accuracy can be caused by the data set itself, so trying more prediction algorithms may not help. In this case, the data set should be re-collected or replaced at this time, or external variables should be integrated to assist prediction.

We also notice that there are some points for future work. The CDR data set is chosen because it represents the mobility of massive people which is also utilized in the original predictability theory of human mobility [17]. But the location information is coarse-grained and is not precise enough in some scenarios. In addition, the predictability theory studied in this paper is based on the prediction of the target with a finite number of position states, without considering the influence of information beyond the historical trajectory.

REFERENCES

[1] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, p. 38, 2014.

[2] J. P. Bagrow, D. Wang, and A.-L. Barabási, "Collective response of human populations to large-scale emergencies," *PLoS ONE*, vol. 6, no. 3, Mar. 2011, Art. no. e17680.

[3] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee, "Quantifying the impact of human mobility on malaria," *Science*, vol. 338, no. 6104, pp. 267–270, Oct. 2012.

[4] D. Buhalis and A. Amaranggana, "Smart tourism destinations," in *Information and Communication Technologies in Tourism*, Z. Xiang and I. Tussyadiah, Eds. Cham, Switzerland: Springer, 2013, pp. 553–564.

[5] L. Yao, Q. Z. Sheng, X. Wang, W. E. Zhang, and Y. Qin, "Collaborative location recommendation by integrating multi-dimensional contextual information," *ACM Trans. Internet Technol. (TOIT)*, vol. 18, no. 3, p. 32, 2018.

[6] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 186–194.

[7] S. Jiang, J. Ferreira, and M. C. González, "Clustering daily patterns of human activities in the city," *Data Mining Knowl. Discovery*, vol. 25, no. 3, pp. 478–510, Nov. 2012.

[8] R. Di Taranto, S. Muppirisetty, R. Raulefs, D. Slock, T. Svensson, and H. Wymeersch, "Location-aware communications for 5G networks: How location information can improve scalability, latency, and robustness of 5G," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 102–112, Nov. 2014.

[9] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Next place prediction using mobility Markov chains," in *Proc. 1st Workshop Meas., Privacy, Mobility*, 2012, p. 3.

[10] J. Tkacik and P. Kordik, "Neural turing machine for sequential learning of human mobility patterns," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 2790–2797.

[11] F. Alhasoun, M. Alhazzani, F. Aleissa, R. Alnasser, and F. González, "City scale next place prediction from sparse data through similar strangers," in *Proc. ACM KDD Workshop*, Halifax, NS, Canada, vol. 14, 2017, pp. 191–196.

[12] Q. Lv, Y. Qiao, N. Ansari, J. Liu, and J. Yang, "Big data driven hidden Markov model based individual mobility prediction at points of interest," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5204–5216, Jun. 2017.

[13] L. Liu, S. Zhang, W. Zhou, W. Cai, and Q. Cui, "Diffusion kernel based mobility prediction for wireless users," in *Proc. IEEE Intl Conf Dependable, Autonomic Secure Comput.*, Aug. 2019, pp. 872–875.

[14] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1082–1090.

[15] H. Gao, J. Tang, and H. Liu, "Mobile location prediction in spatio-temporal context," *Nokia Mobile Data Challenge Workshop*, vol. 41, no. 2, pp. 1–4, Jun. 2012.

[16] Q. Lv, Y. Di, Y. Qiao, Z. Lei, and C. Dong, "Spatial and temporal mobility analysis in LTE mobile network," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2015, pp. 795–800.

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

[18] S.-M. Qin, H. Verkasalo, M. Mohtaschemi, T. Hartonen, and M. Alava, "Patterns, entropy, and predictability of human mobility and life," *PLoS ONE*, vol. 7, no. 12, Dec. 2012, Art. no. e51353.

[19] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Sci. Rep.*, vol. 3, no. 1, Dec. 2013, Art. no. 2923.

[20] V. Kulkarni, A. Mahalunkar, B. Garbinato, and J. Kelleher, "Examining the limits of predictability of human mobility," *Entropy*, vol. 21, no. 4, p. 432, Apr. 2019.

[21] Y. Liao and S. Yeh, "Predictability in human mobility based on Geographical-boundary-free and long-time social media data," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 2068–2073.

[22] Y. Han, W. Sun, and B. Zheng, "Ineffectiveness of dictionary coding to infer predictability limits of human mobility," 2018, *arXiv:1810.06405*. [Online]. Available: http://arxiv.org/abs/1810.06405

[23] P. Xu, L. Yin, Z. Yue, and T. Zhou, "On predictability of time series," *Phys. A, Stat. Mech. Appl.*, vol. 523, pp. 345–351, 2019.

[24] H.-T. Zhang, T. Zhu, D. Fu, B. Xu, X.-P. Han, and D. Chen, "Spatiotemporal property and predictability of large-scale human mobility," *Phys. A, Stat. Mech. Appl.*, vol. 495, pp. 40–48, Apr. 2018.

[25] A. Cuttone, S. Lehmann, and M. C. González, "Understanding predictability and exploration in human mobility," *EPJ Data Sci.*, vol. 7, no. 1, pp. 1–17, Dec. 2018.

[26] J. Guo, L. Liu, S. Zhang, and J. Zhu, ''Mobility prediction with missing locations based on modified Markov model for wireless users,'' in *Proc. IEEE Int. Congr. Big Data (BigDataCongress)*, Jul. 2019, pp. 132–138.

[27] L. Liu, W. Zhou, S. Zhang, and W. Cai, ''Composite Gaussian function modeling of mobility prediction accuracy for wireless users,'' in *Proc. IEEE Intl Conf Dependable*, Aug. 2019, pp. 761–767.

[28] L. Liu, S. Zhang, and W. Zhou, ''Mobility predictability of college students via full lifecycle campus consuming logs,'' in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[29] M. Lin, W.-J. Hsu, and Z. Q. Lee, ''Predictability of individuals' mobility with high-resolution positioning data,'' in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 381–390.

[30] W. Chen, Q. Gao, and H. Xiong, ''Temporal predictability of online behavior in foursquare,'' *Entropy*, vol. 18, no. 8, p. 296, Aug. 2016.

[31] P. Baumann and S. Santini, ''On the use of instantaneous entropy to measure the momentary predictability of human mobility,'' in *Proc. IEEE 14th Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2013, pp. 535–539.

[32] G. Smith, R. Wieser, J. Goulding, and D. Barrack, ''A refined limit on the predictability of human mobility,'' in *Proc. IEEE Int. Conf. Pervas. Comput. Commun. (PerCom)*, Mar. 2014, pp. 88–94.

[33] A. Grigoryan, *Heat Kernel and Analysis on Manifolds*, vol. 47. Providence, RI, USA: American Mathematical Society, 2009.

[34] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.

[35] 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 Trans. Inf. Theory*, vol. 44, no. 3, pp. 1319–1327, May 1998.

[36] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, ''Large-scale mobile traffic analysis: A survey,'' *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 124–161, 1st Quart., 2016.

[37] T. H. Rashidi, A. Abbasi, M. Maghrebi, S. Hasan, and T. S. Waller, ''Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges,'' *Transp. Res. C, Emerg. Technol.*, vol. 75, pp. 197–211, Feb. 2017.

[38] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, ''On the levy-walk nature of human mobility,'' *IEEE/ACM Trans. Netw.*, vol. 19, no. 3, pp. 630–643, Jun. 2011.

[39] J. (. Xia, P. Zeephongsekul, and C. Arrowsmith, ''Modelling spatio-temporal movement of tourists using finite Markov chains,'' *Math. Comput. Simul.*, vol. 79, no. 5, pp. 1544–1553, Jan. 2009.

[40] H. Nawaz, H. M. Ali, and S. U. R. Massan, ''A study of mobility models for UAV communication networks,'' in *Proc. Tecnología_Glosas de Innov. Appl. Pyme*, May 2019, pp. 276–297.

JUNYAO GUO received the B.E. degree in electronics and information engineering from the Southwest University of Science and Technology (SWUST), China, in 2018. She is currently pursuing the M.E. degree with the College of Information Science and Technology, University of Science and Technology of China (USTC), Hefei, China. Her research interests include data mining, artificial intelligence, mobility prediction, and intelligent communications.

SIHAI ZHANG (Senior Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, China, in 2006.

He has been with the Key Laboratory of Wireless-Optical Communications, University of Science and Technology of China. He is currently an Associate Professor at the Department of Electronic Engineering and Information Science, USTC. He has authored or coauthored over 60 technical articles, such as the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, MONET, and WPC. He initiated the research field of wireless big data, in 2014. His research interests include wireless communication and networks, big data analysis, and machine learning. He has participated in many projects, including the National Science Foundation of China for Machine Type Communications and the Key Program of the National Natural Science Foundation of China for Wireless Big Data. He has been co-organizing five annual workshop on Wireless Big Data, China, since 2014. He has served over 15 international conferences, as a member of organizing committee, a TPC member or a reviewer, such as the Publication Chair for WCSP 2014. In 2016, he Co-Chaired special sessions on Wireless Big Data in WCSP 2016 and Machine Type Communications in WPMC2016. He has been a Guest Editor of Special Issue on Wireless Big Data of JCIN and on Recent Advances in Ultra Dense 5G Networks with Application to Machine Type Communication of IJDSN, in 2017.

JINKANG ZHU (Life Member, IEEE) received the B.S. degree in electrical engineering from Sichuan University, China, in 1966. He joined the University of Science and Technology of China (USTC), in 1966, where he has been a Professor, since 1992. He has been committed to research on wireless mobile communications and networks, signal processing for communications, and the future wireless technologies.

RUI NI (Member, IEEE) received the B.E. and Ph.D. degrees in electrical and electronics engineering from the University of Science and Technology of China (USTC), in 2006 and 2011, respectively. Since 2011, he has been the Principal Engineer with the Wireless Technology Laboratory (WTLAB), 2012 Laboratory, Huawei Technologies Company Ltd. His research involves various architectures of radio access network and core network from 2G to 5G, especially network slicing in the 5G era. He has a rich wireless network related experience in engineering practice. His current interests include orbital angular momentum, large intelligence surface, and wireless big data processing for beyond 5G systems.