Predictability of Individuals' Mobility with High-Resolution Positioning Data

Miao Lin
Nanyang Technological University
linm0018@e.ntu.edu.sg

Wen-Jing Hsu
Nanyang Technological University
hsu@pmail.ntu.edu.sg

Zhuo Qi Lee
Nanyang Technological University
ZQLEE1@e.ntu.edu.sg

ABSTRACT

The ability to foresee the next moves of a user is crucial to ubiquitous computing. Disregarding major differences in individuals' routines, recent ground-breaking analysis on mobile phone data suggests high predictability in mobility. By nature, however, mobile phone data offer very low spatial and temporal resolutions. It remains largely unknown how the predictability changes with respect to different spatial/temporal scales. Using high-resolution GPS data, this paper investigates the scaling effects on predictability. Given specified spatial-temporal scales, recorded trajectories are encoded into long strings of distinct locations, and several information-theoretic measures of predictability are derived. Somewhat surprisingly, high predictability is still present at very high spatial/temporal resolutions. Moreover, the predictability is independent of the overall mobility area covered. This suggests highly regular mobility behaviors. Moreover, by varying the scales over a wide range, an invariance is observed which suggests that certain trade-offs between the predicting accuracy and spatial-temporal resolution are unavoidable. As many applications in ubiquitous computing concern mobility, these findings should have direct implications.

Author Keywords

predictability; entropy; mobility management;

ACM Classification Keywords

H.2.8 Data Management: Spatial databases and GIS; I.5.2 Pattern Recognition: Design methodology

General Terms

Algorithms, Experimentation, Human Factors, Measurement.

INTRODUCTION

The ability to foresee individuals' future whereabouts is crucial for personal positioning [3], epidemic prevention [11], city planning [7], etc. In mobile and pervasive computing, probabilistic models, such as Markov models [1, 10, 2], Bayes models [13, 6], pattern mining method [9] have been proposed to continuously predict the next move of individuals' mobility. Although the predicting accuracy has been greatly improved because of these efforts, little is known whether the predicting accuracy is already approaching the limit or whether further research efforts will yield diminishing returns. Moreover, the predicting accuracy is apparently affected by the scale of the locations and the time interval concerned. For instance, while it is relatively easy to predict if the required spatial-temporal resolutions of the prediction are low, e.g., the person will probably be in the same country tomorrow, it is challenging to predict exactly where a person will be located between, say, 1045-1100am. How does the prediction accuracy vary with respect to the spatial-temporal scale? Again, little is known beyond intuition level, and it is difficult (or even meaningless) to compare the existing results as such.

In their breakthrough research on human mobility, Song et al. [14] study the predictability issue based on the analysis of mobile phone data over a large population. *Predictability* is defined as the information-theoretic upper bound that fundamentally limits any mobility prediction algorithm in predicting the next locations based on historic records. Disregarding the apparent differences in individuals' daily routines, their analysis revealed a 93% predictability in individuals' mobility. This main finding is fundamental to many related fields[14].

By nature, however, the mobile phone data exhibit low resolution in both spatial and temporal dimensions, and they may be inadequate for further mobility research. In research that uses mobile phone data [14, 5], a user's "location" is identified with the service coverage area of the base station, and the user's locations are captured when he/she initiates or receives a call. Since it is extremely difficult to delineate the exact coverage area of a base station, it is approximated by using a Voronoi diagram, which may give a wrong indication of user's actual location. Moreover, using this scheme, the "locations" are generally of heterogeneous shape, and they differ vastly in size, with the larger "locations" one or two orders of magnitude larger than the smaller ones. At an average size of $3km^2$, a base-station imposed "location" may contain several distinct locations of interest. Furthermore, the number of instances collected from mobile phones is limited by the calls generated/received by the users, which could be as low as 3-4 per day, and hence the hourly location string often contains a significant fraction of unknown loca-

UbiComp '12, Sep 5-Sep 8, 2012, Pittsburgh, USA.

tions[14]. Therefore, it is inherently difficult to further the mobility study based on the phone data, and presently it is unclear whether the predictability findings remain valid with high-resolution data, and how the predictability varies under different spatial-temporal scales, e.g., with smaller locations or over shorter time intervals.

Thanks to the availability of large GPS datasets [15, 16] researchers can begin to address these issues. GPS data offer much higher resolutions. In the data sets, the sampling frequency ranges between a few seconds to less than a minute. The spatial resolution is in the order of 10 meters. Thus GPS data offer two orders of magnitude of improvements in both spatial and temporal resolution. The new data enable us to study how the predictability varies at different scales of the size of locations and the length of time intervals.

The approach for analyzing the predictability is as follows.

- Given the GPS records and the required spatial-temporal resolutions, a grid map is applied to encode the locations, converting the records into a long string of location symbols, based on which the entropy is estimated.
- The time scale is easily specified by a duration. The grid map provides a uniform way to scale the spatial dimension, i.e., to vary the size of a location from that of a building to a village or even much larger.
- The predictability is calculated in relation to the estimated entropy of the string of location symbols.
- Several experiments are conducted based on different sizes of the locations.

Our contributions lie in a few aspects.

- Firstly, our research refines the current knowledge about predictability, and it removes the difficulties arising from assigning a user's location with the mobile phone data. For instance, our analysis reveals that the predictability can be as high as 90% at an hourly temporal resolution and a spatial resolution where the size of each location is roughly that of a large building–or about 1/20 of the average coverage area of cell towers. Moreover, the predictability is found to be independent of the size of the mobility area covered by each individual. Although the mobility area varies greatly among the individuals, a rather consistent predictability is obtained.
- Secondly, by varying the spatial scales of the grid cells over a wide range, we observe an invariance between the predictability and spatial resolution (or spatial uncertainty), namely they can't both achieve high accuracy simultaneously. The invariance suggests that trade-offs may be needed between these two measures when designing algorithms for mobility predictions.
- We also confirm both theoretically and empirically that redundancy and predictability are effectively the same statistical quantity in the context of mobility.

EXISTING WORK

Two papers have pioneered the study of the predictability of human mobility [14, 8]. Song et al. [14] first defined the predictability as the limit of any algorithm for predicting individuals' future locations. The predictability is derived from the estimated entropies of the mobile phone data. The predictability is centered around 93% over a large population, independent of the size of the area covered by individuals' mobility or other demographic factors. As noted earlier, the high predictability is obtained based on low resolution positioning data. Since the service area of each cell tower represents the locations in the paper, the average size of a "location" is roughly 3 km^2. For higher resolution positioning data, it is unknown whether the human mobility is still highly predictable, which motivates the current study. More importantly, the size and shape of the area covered by each cell tower is totally irregular.

Jensen et al. [8] apply the same theory to analyze the predictability based on various types of mobile sensor data of 48 days' records on average for 14 individuals. A similar high predictability is reported based on the discrete time series constructed from GSM, WLAN, blue tooth and acceleration data at a given window length independently. They further introduce the predictive information, representing the mutual information between the entropies with or without knowing the past history. By varying the time scales from a few minutes to a few hours, the highest predictive information is obtained when the time scale is 4 to 5 minutes. In Lempel-Ziv data compression algorithm, a comparison between the states in the discrete time series is needed, but the measurements that quantify the similarity between the states are not presented in the paper. Also, their findings may suffer from the limit of each type of data. For instance, both WLAN and blue tooth data are capable of revealing the mobility of an individual in a small region but not between the regions, thus the high predictability based on these data does not generalize to wide-area mobility. For GSM data, the mobility behaviors over a large region can be observed, but they are of low resolution, and suffer from similar disadvantages with mobile phone data- where the locations are of irregular shape and uneven size.

MOBILITY DATA

The large GPS data sets used in this paper contain the movement records of 40 individuals around China between 2008 and 2009. The 40 individuals' records are a subset of data from the data sets released by Yu Zheng [15, 16]. In the complete data sets, almost half of the individuals are college students, and the remaining are government staff and employees from Microsoft and several other companies. Each individual's trajectory consists of a series of time-stamped positioning readings in pairs of longitude and latitude. For example, the following GPS readings show a partial trajectory of one individual

x_1={(116.3329° 39.9687°), 2008-01-24 13:16:21},
x_2={(116.3331° 39.9674°), 2008-01-24 13:19:18},
x_3={(116.3331° 39.9667°), 2008-01-24 13:20:54},

In the experiments, a 16 week-long trajectory is extracted for each individual. To categorize different mobility behaviors, we define the radius of gyration r_g as in [14]

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(surface\_distance(p_i, p_c))^2} \qquad (1)$$

where p_i represents the location at time i, and $p_c = \frac{1}{n}\sum_{i=1}^{n} p_i$ is the center of the trajectory. $surface\_distance$ denotes a function calculating the minimum distance over the earth's surface between the locations given in longitude and latitude. The radius of gyration of the individuals in the data set ranges from a few kilometers to more than one thousand kilometers.

The location string for each individual is encoded according to the spatial and temporal resolution into a string of location symbols. At an hourly resolution, for instance, the length of each location string for each individual contains $16 \times 7 \times 24 = 2688$ location symbols.

METHODOLOGY

Here we present the framework for analyzing the predictability of human mobility.

Grid cells and locations

DEFINITION 1. A *grid cell* is a rectangular region characterized by two parameters, namely the *origin* $= (long_i, lat_i)$ and *spatial scale*, denoted by s. The origin is a given point of reference. The parameter s is a numeric value used to specify the size of the grid cell. A grid map is a collection of disjoint grid cells that collectively cover the mobility area in a GPS data set.

Strictly speaking, a region on Earth that is bounded by pairs of longitudinal and latitudinal lines is not rectangular. However, this simplified description may fit well when the scale is small.

Suppose the grid cell of a given grid map has a scale s. The increment of adjacent grid points differ by $0.001^\circ s$ in longitude or latitude. Let $p_0 = (long_0, lat_0)$ denote the origin of the grid map. The grid cell on the i-th row and j-th column of the map, labeled as $Cell(i, j)$ is bounded by its four corners, $p_{ij_k}, k = 1, 2, 3, 4$. In terms of longitude and latitude, $p_{ij_1} = (long_0 + 0.001 \times (i-1) \times s, lat_0 + 0.001 \times (j-1) \times s)$, $p_{ij_2} = (long_0 + 0.001 \times i \times s, lat_0 + 0.001 \times (j-1) \times s)$, $p_{ij_3} = (long_0 + 0.001 \times (i-1) \times s, lat_0 + 0.001 \times j \times s)$, $p_{ij_4} = (long_0 + 0.001 \times i \times s, lat_0 + 0.001 \times j \times s)$. Figure 1(a) shows the mapping of one individual's partial trajectory to a grid map, and Figure 1(b) shows the trajectory in terms of the transitions between the grid cells.

The distance between the locations indicated by pairs of longitude and latitude are expressed in kilometers. One degree of longitude or latitude covers different length. For instance, around Beijing, China, 0.001° corresponds to 0.087 km in longitude and 0.111 km in latitude respectively. Thus, the area of each grid cell is controlled according to the scale factor n, and the position of each grid cell can be varied according to the choice of the origin.

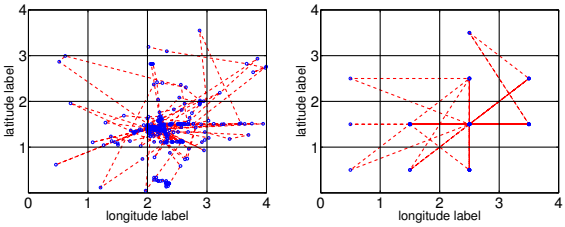


Figure 1. (a) An overview of one individual's partial trajectory on a grid map. (b) The trajectory is converted into a series of transitions between grid cells.

When mapping each individual's trajectory to a grid map, the trajectory is converted to a location string. At a given temporal resolution, there may be multiple distinct locations recorded in a time interval, or there may be no records at all. For multiple locations, as with [14], the location for such a duration is chosen randomly according to the probabilities of visiting these locations recorded during that time interval.

For missing records, we apply the following rule originally from [1]. The GPS signals from the satellites are often disrupted inside a building. Thus, when the ending and beginning locations of a gap in the GPS records are the same, the user is taken as dwelling at the same location during that time. This rule also caters for the situation where the individual enters a building and exits later, or where the individual turns off the GPS devices in an indoor place. Otherwise, the user's location during the time is considered as unknown and is represented by a "?" symbol. How to estimate the entropy from a location string that contains the unknown locations is discussed in the experimental part.

Entropy measurements

Let $\mathbf{X} = (x_1, x_2, ..., x_n)$ be a discrete time location string converted from a raw trajectory, where n denotes the time instance, and each x_i represents a location encoded in terms of the grid cells. Suppose the location string X is generated from a distribution $p(x) = Pr(X = x)$, where x is from a finite set $A = \{C_i\}$. The following information-theoretic definitions of entropy may be applied to X.

DEFINITION 2. The random entropy H^{rand} of X is defined as $H^{rand} = \log N$, where N denotes the number of distinct locations in the location string. And, log is based 2 throughout this paper.

DEFINITION 3. The *temporal-uncorrelated entropy* of X is defined as

$$H(X)^{unc} = -\sum_{x \in A_s} p(x) \log p(x) \qquad (2)$$

The random entropy entirely ignores the frequencies of the locations. Hence it is an over-estimate of the actual randomness of mobility. The second entropy measure takes the frequencies into consideration. The distribution of $p(x)$ can be estimated through a maximum likelihood estimator, where $p(x)_{x\in A}$ is estimated from the current location string. This measure of randomness is temporal-uncorrelated, as it totally ignores the order of location symbols in the string.

DEFINITION 4. The *entropy rate* $H = H(X)$, also called "per-symbol" entropy, of X is defined as

$$H = H(X) = \lim_{n\to\infty} \frac{1}{n} H(X1, X2, ..., Xn) \quad (3)$$

where $H(X1, X2, ..., Xn)$ is the entropy of the joint distribution of random variables $X_1^n = (X1, X2, ..., Xn)$.

The entropy rate takes the order of locations into considerations, which can better measure the predicability of mobility. This measure can be estimated through Lempel-Ziv data compression algorithm [12].

Given a location string $x = (..., x_{-1}, x_0, x_1, ...)$, let x_i^j represent a subsequence beginning from index i and ending at j. L_i^n denotes the shortest substring starting at index i that does not appear in the window x_{i-n}^{i-1} of length n.

$$L_i^n = L_i^n(x_{i-n}^{i+n-1}) = 1 + \max\{0 \le l \le n : x_i^{i+l-1} = x_j^{j+l-1}, i-n \le j \le i-1\} \quad (4)$$

L_i^n is said to be a sliding window of length n. Similarly, L_i^i represents a matching, where each time a match is searched through the entire history. Since the entropy estimator based on a sliding window fluctuates with respect to the window size, we choose an increasing window matching method for estimating the entropy. The result is denoted as $\hat{H}_n$,

$$\hat{H}_n = (\frac{1}{n}\sum_{i=2}^{n} \frac{L_i^i}{\log i})^{-1} \quad (5)$$

Redundancy and predictability

DEFINITION 5. The redundancy of a sequence X, denoted by $\rho(X)$, is defined by

$$\rho(X) = (H^{max}(X) - H^{true}(X))/H^{max}(X), \quad (6)$$

where H^{max} denotes the maximum entropy and H^{true} denotes the true entropy of the sequence. In our case, $H^{max} = H^{rand}$ and $H^{true} = \hat{H}_n$.

The redundancy is equivalent to the compression rate of a sequence, which is not the only factor that depends on the predictability of the sequence [4]. In the following, we show that the redundancy and predictability are roughly equal quantities in the context of mobility sequences.

DEFINITION 6. Let Π^{max} denote the maximum probability based on a given location string with the entropy $\hat{H}_n$, and N the number of distinct locations presented in the location string. $\Pi^{max}(\hat{H}_n, N)$ is the solution from the equation,

$$\hat{H}_n = -[\Pi^{max}\log\Pi^{max} + (1-\Pi^{max})\log(1-\Pi^{max})] + (1-\Pi^{max})\log(N-1) \quad (7)$$

It is known that [14], $\Pi^{max}(\hat{H}_n, N)$ is an upper bound of the predictability, which is achievable by always predicting the historically most probable location as the next location.

LEMMA 1. *The redundancy ρ and predictability Π^{max} are equivalent statistical quantities on measuring mobility sequences.*

Before proving this lemma, we recall the definition of the binary entropy of a random variable.

DEFINITION 7. Let Y denote a binary variable whose value is chosen from two values with probability $p_Y(i)$, where $i = 1, 2$ and $p_Y(1) + p_Y(2) = 1$. The binary entropy of Y, denoted by $H_b(Y)$, is given by

$$H_b(Y) = -[p_Y(1)\log p_Y(1) + p_Y(2)\log p_Y(2)]. \quad (8)$$

Therefore $-[\Pi^{max}\log\Pi^{max} + (1-\Pi^{max})\log(1-\Pi^{max})]$ is the binary entropy, denoted by $H_b(\Pi^{max})$ with slight abuse of the notation, of the variable which indicates the predicting results. Since the binary variable is chosen from two values, with probability Π^{max} the algorithm makes a correct prediction and with the complimentary probability $1-\Pi^{max}$ the algorithm fails to predict the next move. Now we are ready to prove Lemma 1.

Proof According to the definition 7, Equation(7) is rewritten as

$$\hat{H}_n = H_b(\Pi^{max}) + (1-\Pi^{max})\log(N-1) \quad (9)$$

$$\hat{H}_n \approx H_b(\Pi^{max}) + (1-\Pi^{max})\log N \quad (10)$$

$$\hat{H}_n/H^{rand} \approx H_b(\Pi^{max})/H^{rand} + (1-\Pi^{max}) \quad (11)$$

$$1 - \hat{H}_n/H^{rand} \approx \Pi^{max} - H_b(\Pi^{max})/H^{rand} \quad (12)$$

$$\rho \approx \Pi^{max} - H_b(\Pi^{max})/H^{rand} \quad (13)$$

Equation 10 is obtained from Equation 9 by approximating $\log(N-1)$ with $\log N$. Equation 11 is obtained by dividing both sides with $\log N$ and replacing $\log N$ with H^{rand}. Equation 12 is obtained by rearranging the terms in Equation 11. The last equation is obtained from the definition of ρ.

From Equation 12, we can see that redundancy and predictability differ by the ratio the two measures of entropy, i.e., $H_b(\Pi^{max})/H^{rand}$. Since the binary entropy $H_b(\Pi^{max})$ is within the range of 0 to 1, $H_b(\Pi^{max})$ is small, especially when Π^{max} is close to 1. The random entropy H^{rand} depends on the number of distinct locations, which is more than 64 when the spatial scale is 18, and more than 128 when the scale is 4. Thus the entropy ratio is minor in mobility data.

Consequently, the redundancy and predictability may be con-

sidered as the same quantity when measuring the statistical property of mobility sequences, which implies that high redundancy equals to high predictability and vice versa. □

In the experiments, we will verify quantitatively this relation between predictability and the redundancy with the mobility sequences.

EXPERIMENTAL RESULTS

In this section, we present the experimental results. The following sets of experiments are carried out:

1. Refining the method for estimating the entropy from incomplete mobility sequences.
2. Evaluating the impacts of varying the spatial scale and the temporal scale on the entropy measures.
3. The entropy measures and predictability for 40 individuals at two spatial scales.
4. The relationship between the predictability and spatial uncertainty (to be explained later) by varying the scales of the locations.
5. How the predictability varies with respect to the radius of gyration.
6. The effect of choosing different origins of the grid map on predictability.

From the experiments, we found that the entropy measures are more influenced by the spatial scales, rather than the temporal scales. In the subsequent experiments, we choose to vary only the spatial scale. When the spatial scale is 18, the size of each grid cell is around 3 km^2, which is similar to the service area of a cell tower in [14]. When the scale is 4, the size of each grid cell is approximately 0.15km^2, or that of a large building.

Estimating entropy from incomplete mobility sequences

In the data sets, none of the individuals has complete records. Figure 2 shows the fraction of unknown locations when the spatial scale is 4 and 18 respectively. When the scale is large (or, alternatively, when the grid cells are large), the fraction of unknown locations is smaller since more data collected in a larger area are mapped into a grid cell, making it more likely to encounter the same grid cell label. For the majority of the individuals, the fraction of unknown location is between 0.2 to 0.9, which is much lower than the fraction of unknown locations between 0.7 to 0.9 in the mobile phone data [14].

In [14], after analyzing 8-day complete data, it was concluded that the estimated entropy may be considered as the true entropy when the fraction of unknown locations is less than 25%. They also show that the errors from the estimated entropy decreases when the length of the location string increases. In our case, location strings are 16-week long and the fraction of unknown location in them is less than 25%. As such, the estimated entropy may be considered as reflecting the true entropy.

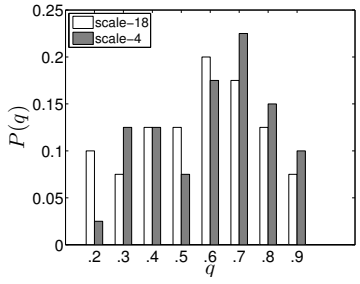


Figure 2. The distribution of the fraction of unknown locations.

Using mobile phone data, Song et al. [14] discovered a linear relation between the fraction of unknown locations, denoted as q, and the ratio of entropy calculated from the partial string and a randomly shuffled string, denoted as $\ln(H(q)/H^r(q))$. Therefore, the estimated entropy is multiplied by a ratio inferred at $q = 0$. With GPS data, although a similar linear relationship can be observed for a few individuals, the ratio $\ln(H(q)/H^r(q))$ varies greatly when $q = 0$. Thus, we are forced to look for a new method for estimating the entropy from a partial location string, which is described in the following.

We have rather complete data with a few individuals, and the ratio of missing data $q < 0.25$. Suppose the true entropy is H for a location string. For a string with q fraction of unknown locations, an estimated entropy, denoted as $\hat{H}(q)$, is calculated by using the Lempel-Ziv data compression algorithm given by Equation(5). The following steps are carried out based on these individuals' data to explore the relation between the estimated entropy and the fraction of unknown locations.

- Assign $H = \hat{H}(q)$ (since $q < 0.25$)
- Randomly replace Δq fraction of locations by "?", in order to mimic the situation where $q + \Delta q$ fraction of locations are unknown.
- The estimated entropy, denoted as $\hat{H}(q + \Delta q)$, is calculated from the substring by removing all the "?".
- Let Δe be the error of the estimated entropy, calculated by $\Delta e = \frac{\hat{H}(q+\Delta q) - \hat{H}(q)}{\hat{H}(q)}\%$.
- By gradually increasing Δq, we can generate a plot for the error Δe versus the fraction of unknown locations $(q + \Delta q)$.

By using the preceding procedure, we are able to analyze how the error of the estimated entropy varies with respect to the fraction of unknown locations. Figure 3(a) and Figure 3(c) show rather consistent error curves independent of the individuals and the scales. Based on this, we calculate the mean errors across different individuals at a given spatial scale, and a linear curve is fitted, as shown in Figure 3(b) and Figure 3(d). For the individuals with a greater fraction of unknown locations (than 0.25), the error for the estimated entropy is inferred from the fitted curve.

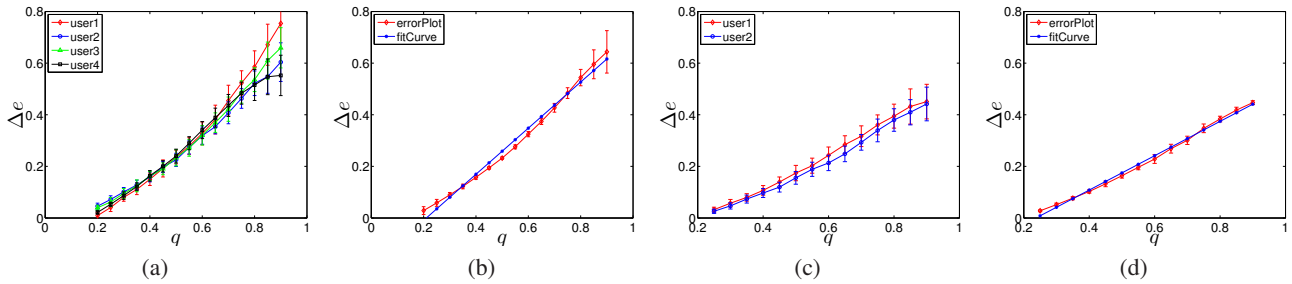


Figure 3. (a) At the spatial scale 18, there are 4 individuals whose fraction of unknown locations is less than 0.25. (b) When applying the linear curve $y = p_1 \times x + p_2$ to fit the mean error curve, the parameters are $(p_1, p_2) = (0.8932, -0.1878)$. For each individual, two error curves are obtained by changing the origins twice for the grid map. The first origin is chosen from the minimum value of longitude and latitude in the person's historical records, and the second origin is randomly chosen from his trajectory. (c) At scale 4, there are 2 individuals whose fraction of unknown locations is less than 0.25. (d) A linear curve is used to fit the error plot under scale 4, which has the parameters: $(p_1, p_2) = (0.6655, -0.158)$.

Predictability vs spatial-temporal resolutions

To evaluate the effects of spatial and temporal scales (sampling rate) on predictability, two individuals with more complete records are chosen. We define the blackout periods, between 8 pm and 6 am of the following day. During the blackout periods, the missing locations are treated the same as the previous locations recorded. The two individuals exhibit significantly different mobility behaviors, with their radius of gyration being 325 kilometers and 15 kilometers, respectively. Figure 4 and Figure 5 show the effects of spatial and temporal scales on H^{rand}, $\hat{H}_n$, and Π^{max}.

- Figure 4 and Figure 5 show that the higher temporal precision is required, the lower predictability can be achieved; moreover, the effect of the temporal scales on H^{rand}, $\hat{H}_n$ and Π_{max} is clearly linear.
- For the first individual, the spatial scale have more effects, since the values at different spatial scales are well separated from each other. By comparison, for the second individual, the changes in $\hat{H}_n$ and Π^{max} are minor at different spatial scales. With a radius of gyration of more than 300 kilometers, the first individual exhibits a rather diversified mobility behavior, while the second individual has a radius of gyration of around 15 kilometers.

Thus the experiments suggest that **changing temporal scales has similar effects on the predictability of different individuals, while changing the spatial scale has different effects, depending on the mobility characteristics of each individual.**

Also, larger temporal scale leads to higher predictability, independent of the spatial scale chosen, as it is more likely to sample the same locations, which increases the redundancy of the mobility sequence and hence the predictability.

Entropy and predictability

Figure 6(a) and Figure 6(b) show the three entropy measures for the spatial scales 18 and 4 respectively. For both spatial scales, the values of the three entropy measures are well-separated, with the random entropy covering the widest range and the entropy rate narrowly distributed. For the largeer grid cells (scale 18), the random entropy varies between 4 and 8, suggesting that the number of distinct locations visited by each individual varies between $2^4 = 16$ and $2^8 = 256$. Similarly, for the smaller grid cells (scale 4), the number of distinct locations varies between $2^5 = 32$ and $2^9 = 512$.

In comparison, the entropy rate ranges between 0.5 and 2 in both cases, suggesting that the "effective" number of locations for each individual is rather consistent, varying from $2^{0.5} = 1.4$ to $2^2 = 4$. **Our study based on the high spatial resolution GPS data confirms the conjecture that although occasionally the individuals may visit a large number of distinct locations, most of the time they merely revisit a few locations, such as home and work place.**

Interestingly, for the larger grid cells (scale 18– where each grid cell is about the average size of the locations in mobile phone data), the range of entropy rate and random entropy are similar to that reported in [14]. When we further shrink the size of each grid cell to roughly 1/20 of the size, we observe an increase in the entropy value.

Figure 6(c) shows the histograms of predictability, with a fitted curve for each plot. When the scale is 18, the predictability is peaked at 93% for all the individuals. The strong law of large number guarantees that when the sample is large enough the distribution for the value of predictability follows a normal distribution. Thus, a normal distribution is used to fit each plot. A Kolmogorov-Smirnov Test accepts the hypothesis that the distribution of the predictability follows a normal distribution. Moreover, we observe that when the scale is 4, where each grid covers an area of 0.15 km^2, the overall predictability decreases to 90%. In other words, the number of "effective" locations increases by $2^{0.5} = 1.4$. In words, **the predictability only decreases slightly even when requiring a much higher precision in the spatial dimension.** It should be interesting that human mobility is highly regular even when the size of the locations are similar to that of a large building. This result is an important refinement to the findings reported in [14].

Although the overall trait for the predictability has minor variations (3%) when shrinking the size of the grid cells, for two particular individuals, the predictability decreases

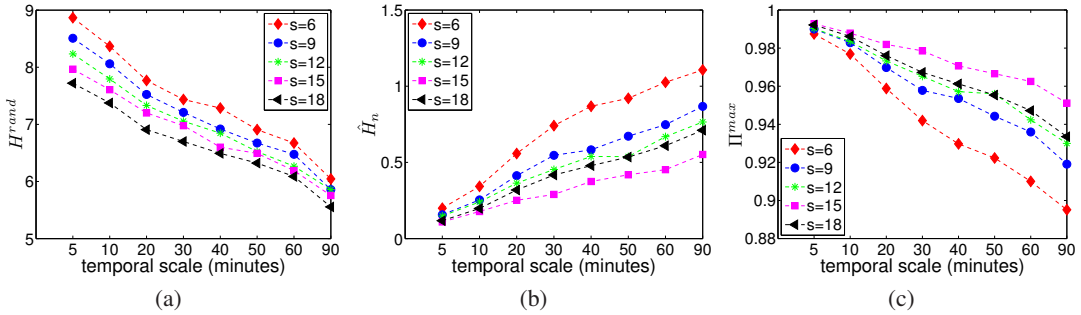


Figure 4. The entropy measures under different spatial-temporal scales: (a) random entropy, (b) estimated entropy rate, and (c) predictability. The values are shown based on the different temporal scales ranging from 5 minutes to 90 minutes. The spatial scale ranges from 6 to 18. This individual exhibits a radius of gyration of about 325 kilometers.

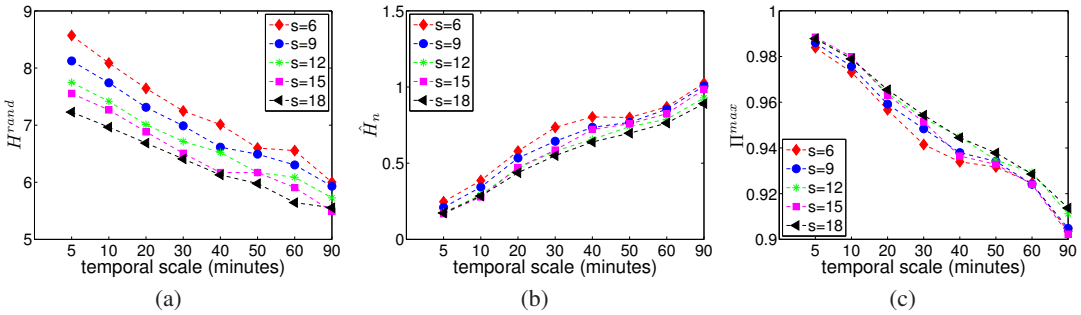


Figure 5. The entropy measures and predictability under different spatial-temporal scales: (a) random entropy, (b) estimated entropy rate, and (c) predictability. The values are shown based on the different temporal scales ranging from 5 minutes to 90 minutes. The spatial scale ranges from 6 to 18. This individual exhibits a radius of gyration of about 15 kilometers.

by about 7%. In detail, one decreases from 94% to 87%, the other decreases from 92% to 85%. The differences trigger the following experiment, which aims to study how the predictability varies with the spatial scale by tuning the scale over a wide range.

Predictability and redundancy

We verify the correlations between predictability and redundancy for all the individuals. Figure 7(a) and Figure 7(b) show that for both scales 18 and 4, the redundancy vary consistently with predictability, which verifies the theoretical derivation, and confirms that the redundancy of each individual's mobility sequence leads to high predictability.

Figure 7(c) indicates that when greatly decreasing the size of locations the redundancy in the individual' mobility sequence also decreases. While reducing the spatial scale can offer better spatial resolution in predictions, it also lowers the redundancy in the mobility sequence, making the prediction of the next symbol more difficult.

Also, the residual between the predictability and redundancy $R_z = H_b(\Pi^{max})/H^{rand}$ is plotted in Figure 7(d), where the individuals are sorted according to the radius of gyration which ranges between a few kilometers to more than 1000 kilometer. **The residual is found to be independent of the radius of gyration.**

Impacts of spatial scale on predictability

This section analyzes how the number of locations and predictability change with respect to the spatial scale, which is varied between 0.1 and 3000. 8 individuals with relatively more complete data are selected, whose fraction of unknown locations is less than 0.4 even when the scale is 0.1. The 8 individuals exhibit rather different mobility behaviors. For 4 individuals, the radius of gyration r_g is more than 200 km, whereas for the rest, r_g is less than 50 km. Under each spatial scale, both the number of locations and predictability are averaged based on 100 randomly chosen origins for the grid maps in order to remove the effects for the choice of origins.

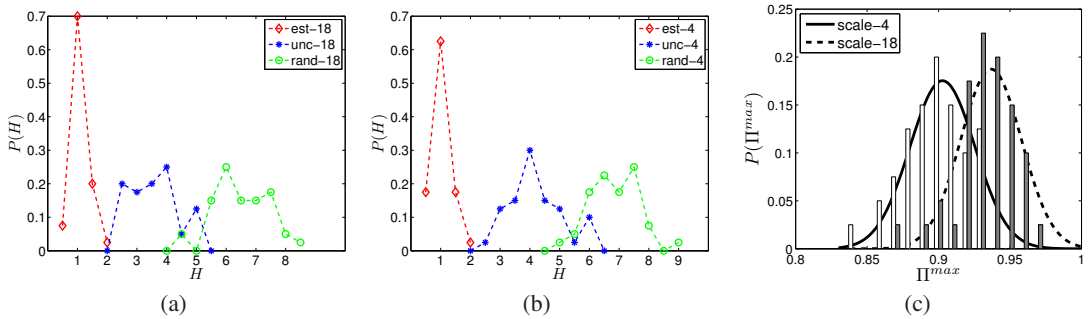
Figure 8(a) indicates a power-law relation for the normalized location ratios when the scale increases for all the individuals. Thus,

$$N_s/N_0 \approx \beta s^{\alpha} \tag{14}$$

Therefore $N_s \sim s^{\alpha}$, where $\alpha = -0.54 \pm 0.02$. This observation reveals certain interesting properties of individual mobility behaviors. An individual's mobility merely covers a small portion of the grid map covering his trajectories. If most of the grid cells had been covered by the trajectory, the rate of decrease would have been close to s^{-2}.

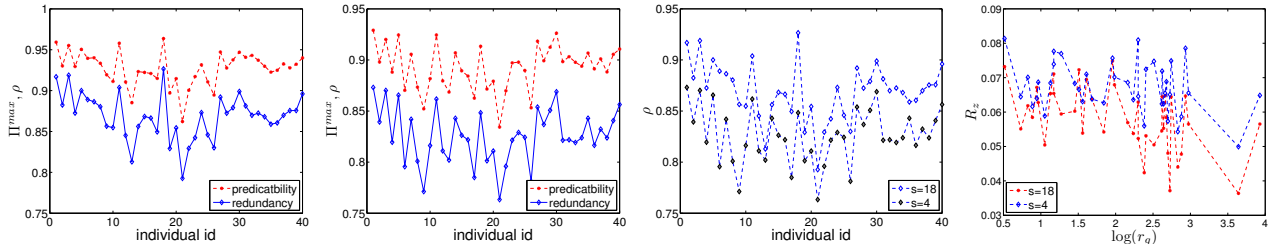
Figure 8(b) shows an important trait for the predictability when increasing the spatial scale.

- Firstly, when $s = 0.1$, which corresponds to about 10^{-4}



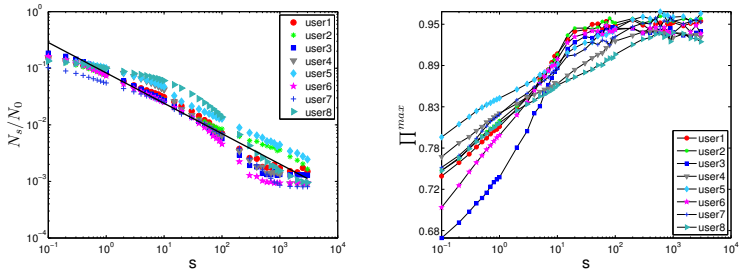
(a) (b) (c)

Figure 6. (a) and (b) show the distributions of the three entropy measures when the spatial scale is 18 and 4, respectively. (c)The histogram for the predictability of all the individuals. The estimated parameters for two normal distributions are (0.903,0.0228) and (0.937,0.021) respectively, where the values in each pair represent the mean and standard deviation of the distribution. A Kolmogorov-Smirnov test is applied to validate whether the normal distribution fits the curve. For the scale 18, the test accepts the null hypothesis at the 5% level of significance, with ks value of 0.1357 and p-value of 0.06. For the scale 4, the test rejects the null hypothesis given ks value is 0.15, p-value is 0.0235. Although the hypothesis is rejected, the plot shows clearly that the distribution peaked at around 90% .



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

Figure 7. The correlation between the redundancy and predictability of human mobility sequence. (a) and (b) show the redundancy and predictability for all the individuals when the scale is 18 and 4, respectively. (c) shows the redundancy for all the individuals for the scales 18 and 4 respectively. (d) The residual among all the individuals (sorted according to the radius of gyration).



(a) (b)

Figure 8. (a) The normalized location ratio versus scale. N_s refers to the number of distinct locations obtained under scale s. N_0 refers to the initial number of GPS readings from each individual's trajectory. The parameters for the fitted line $y = \beta x^{\alpha}$ is $\alpha = -0.54 \pm 0.02$, $\beta = e^{-2.49 \pm 0.11}$. (b) The predictability versus scale. A linear curve is applied to fit the first part of the curve for each individual when the scale is between 0.1 to 100 (the fitted curve is not presented in this figure). The linear curve indicates a power law relationship between Π^{max} and s, where $\Pi^{max} = cs^{\alpha}$ and c is a constant. Among all the individuals, $\alpha \in [0.02, 0.05]$, and $c \in [e^{-0.28}, e^{-.017}]$. Note that this is a log-log plot, where the Y axis is equally spaced according to the exponent (of 10).

km^2 for each location, the predictability for each individual is around 70%, which is much lower than 93% predictability from mobile phone data.

- Secondly, a rather linear relation between the scale and predictability is observed from the log-log plot when $s \in [0.1, 100]$.
- The predictability stays largely constant beyond $s = 100$. In our method, the number of distinct locations is constrained by both the scale and origin. Thus even when the size of a location can cover the whole trajectory of one individual, e.g. when $s = 3000$, the number of distinct location may be more than 1, therefore the predictability can't reach exactly 100%.

From the first part of the plot, we have

$$\Pi^{max} \approx cs^{\alpha} \tag{15}$$

where $s \in [0.1, 100]$ and c is a constant.

The parameter s corresponds to the spatial resolution the algorithm achieves at each predicting step. Let $\mu = \frac{1}{s}$ be the level of spatial uncertainty of the predicting results for the algorithm, where larger μ means greater uncertainty. Therefore,

$$\mu^{\alpha}\Pi^{max} \approx c \tag{16}$$

This implies that **the ability to foresee the future location of an individual (predictability) does not commute with the spatial uncertainty associated with each prediction**. This phenomenon is reminiscent of the uncertainty principle in quantum mechanics where one can't simultaneously determine the exact value of the position and momentum of a particle. In the context of human mobility, the invariance implies a trade-off between the predictability and spatial uncertainty, since it is impossible to design a predicting algorithm using GPS data to achieve both simultaneously with high accuracy.

Predictability vs radius of gyration

Figure 9(a) shows the distribution of the radius of gyration for the 40 individuals, varying between a few kilometers to more than 1000 kilometers. Approximately, 50% of the individuals exhibit a radius of gyration less than 100 kilometers, and the remaining are more than 100 kilometers. The wide range of radius of gyration indicates a great diversity of the mobility behaviors observed from GPS dataset, which help ensure that our experiments are conducted under a general case. When plotting the predictability against $log_{10}(r_g)$, Figure 9(b) shows that the predictability to be independent of the radius of gyration under both scales, which suggests that a large area of mobility does not necessarily lead to a low predictability. When the radius of gyration increases, the predictability does not converge as was reported in [14], as it fluctuates at both spatial scales.

Sensitivity test on the choices of origins

In order to check the effect of the origin of the grid map on the entropy measures, a sensitivity test is conducted. Figure 10(a) and Figure 10(b) show that, the number of distinct locations remains rather invariant when shifting the origin. In rare (2 out of 100) cases, changing the origin does generate a drastically different number of distinct locations. These are considered as outliers. For the majority, (6 out of 8 when the spatial scale is 4, and 7 out of 8 when spatial scale is 18), changing the origin leads to rather constant values for the estimated entropy rate, and no outlier values are observed. For the cases where abnormalities arise, the final results can be smoothed by randomizing the position of the origin for sufficiently large number of times, e.g. 100 times. Thus the shifting of the origins of the grid map has little effect on the results for the number of distinct locations and the estimated entropy rate. Since the predictability is calculated based on the value of these two, it is also considered to be consistent with regard to different origins.

CONCLUSION

The difficulty of identifying a person's location based on mobile phone records has prevented further study on the mobility issues. By using open GPS data sets, this paper refined our understanding about predictability, which is foundational to many ubiquitous applications. On the positive side, our study shows that the predictability can be as high as 90% when a location is about the size of a large building. Also, the study revealed an invariance between the predictability and uncertainty which that trade-offs are needed when designing predicting algorithms. Since the main purpose of this paper is to offer a general upper bound on predicting the next moves of an individual, no algorithm is explicitly given in this paper, which is left for future study.

Because of its fundamental importance, issues related to the predictability of human mobility deserve further study. For lack of sufficiently long and detailed data (except for 2 individuals), this study evaluated the mobility sequences mainly at an hourly sampling rate. How would predictability change at fine-grained temporal scale? Answering this question will require more detailed data to become available. Also, when applying LZ algorithm to estimate the entropy of mobility sequence, the stationarity of the sequence should be satisfied, otherwise the estimated result will have a low convergency rate. Is the mobility sequence for each individual stationary? Our preliminary study based on data from a few individuals suggests that the answer is no. Given this, how can we model mobility for prediction at given temporal-spatial scales? Answering these basic questions will be most useful to the mobility research.

REFERENCES

1. D. Ashbrook and T. Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal Ubiquitous Comput.*, 7(5):275–286, 2003.
2. A. Bhattacharya and S. K. Das. Lezi-update: an information-theoretic approach to track mobile users in pcs networks. MobiCom, pages 1–12, 1999.
3. H. Fang, W.-J. Hsu, and L. Rudolph. Cognitive personal positioning based on activity map and adaptive

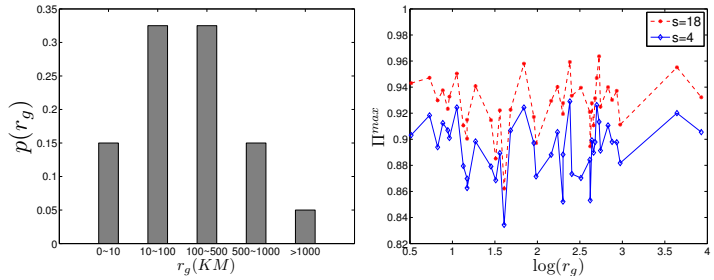


Figure 9. (a) The distribution for the radius of gyration. (b) The plot for the predictability versus radius of gyration.

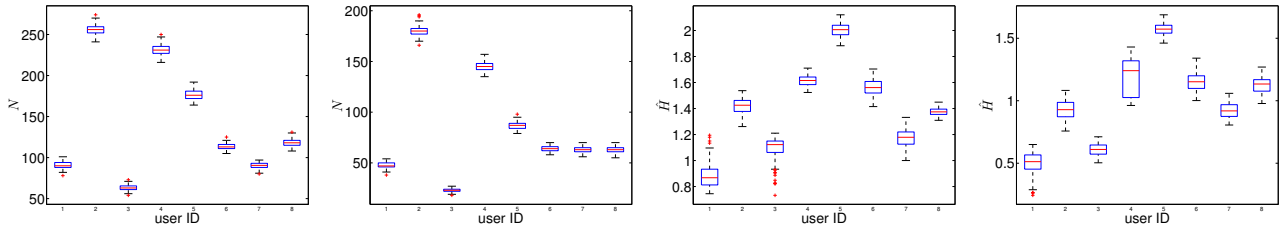


Figure 10. The box plot indicates the median (mark with the box), 25^{th}, and 75^{th} percentiles (the edges of the box), the extremal data points (the whiskers), and outliers (the $+$ sign in red), for both the number of distinct locations and estimated entropy from 8 randomly chosen individuals. (a) and (b) show how the number of distinct locations vary with respect to different origins when spatial scale is 4 and 18 respectively. Similarly, (c) and (d) show how the estimated entropy vary with respect to the different origins for the two spatial scales. For each individual, 100 different origins are generated when encoding the trajectories by a grid map. For the first experiment, the origin is chosen as the minimum value of longitude and latitude from the trajectories, and the origins of the 99 remaining experiments are randomly chosen from the trajectory of each individual.

particle filter. In *MSWiM*, pages 405–412, 2009.

4. M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *Information Theory, IEEE Transactions on*, 38(4):1258 –1270, 1992.

5. M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

6. R. Hariharan and K. Toyama. Project lachesis: parsing and modeling location histories. In *GIS*, pages 106–124, 2004.

7. M. W. Horner and M. E. O'Kelly. Embedding economies of scale concepts for hub network design. *Journal of Transport Geography*, 9(4):255–265, 2001.

8. B. Jensen, J. Larsen, K. Jensen, J. Larsen, and L. Hansen. Estimating human predictability from mobile sensor data. In *MLSP*, pages 196 –201, 2010.

9. H. Jeung, Q. Liu, H. T. Shen, and X. Zhou. A hybrid prediction model for moving objects. ICDE, pages 70–79, 2008.

10. H. Jeung, H. T. Shen, and X. Zhou. Mining trajectory patterns using hidden markov models. DaWak, pages 470–480, 2007.

11. J. Kleinberg. The wireless epidemic. *Nature*, 449:287–288, 2007.

12. I. Kontoyiannis, P. Algoet, Y. Suhov, and A. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *Information Theory, IEEE Transactions on*, 44(3):1319–1327, 1998.

13. J. Krumm and E. Horvitz. Predestination: Inferring destinations from partial trajectories. In *UbiComp*, pages 243–260, 2006.

14. C. Song, Z. Qu, N. Blumm, and A.-L. Barabái. Limits of predictability in human mobility. *Science*, 327:1018–1021, 2010.

15. Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. In *Proc. UbiComp*, pages 312–321, 2008.

16. Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proc. WWW*, pages 791–800, 2009.