www.sciencemag.org/cgi/content/full/327/5968/1018/DC1

Supporting Online Material for

Limits of Predictability in Human Mobility

Chaoming Song, Zehui Qu, Nicholas Blumm, Albert-László Barabási*

*To whom correspondence should be addressed. E-mail: alb@neu.edu

Published 19 February 2010, *Science* **327**, 1018 (2010)
DOI: 10.1126/science.1177170

This PDF file includes

Materials and Methods
SOM Text
Figs. S1 to S13
References

Limits of Predictability of Human Mobility

Supplementary Material

Chaoming Song, Zehui Qu, Nicholas Blumm, Albert-László Barabási

Contents

S1. Data Collection 2

S2. Characterizing Individual Call/Motion Activity 2

S3. Data Preprocessing 4

S4. Determination of User Entropy 4

S5. Fundamental Limits of Predictability 10

S6. Regularity on Weekdays and Weekends 16

S7. The Demographic Dependence 17

References 20

S1. DATA COLLECTION

A. Dataset D_1: This anonymized data set represents 14 weeks of call patterns from 10 million mobile phone users (roughly April through June 2007). The data contains the routing tower location each time a user initiates or receives a call or text message. From this information, a user's trajectory may be reconstructed. For each user i we define the calling frequency f_i as the average number of calls per hour, and the number of locations N_i as the number of distinct towers visited during the three month period.

In order to improve the quality of trajectory reconstruction, we selected 50,000 users with $f_i \geq 0.5$ calls/hour and $N_i > 2$.

B. Dataset D2: Mobile services such as pollen and traffic forecasts rely on the approximate knowledge of customer's location at all times. For customers voluntarily enrolled in such services, the date, time and the closest tower coordinates are recorded on a regular basis, independent of phone usage. We were provided with the anonymized records of 1,000 such users, from which we selected 100 users whose coordinates were recorded every hour over eight 8 days.

S2. CHARACTERIZING INDIVIDUAL CALL/MOTION ACTIVITY

A. Number of visited locations

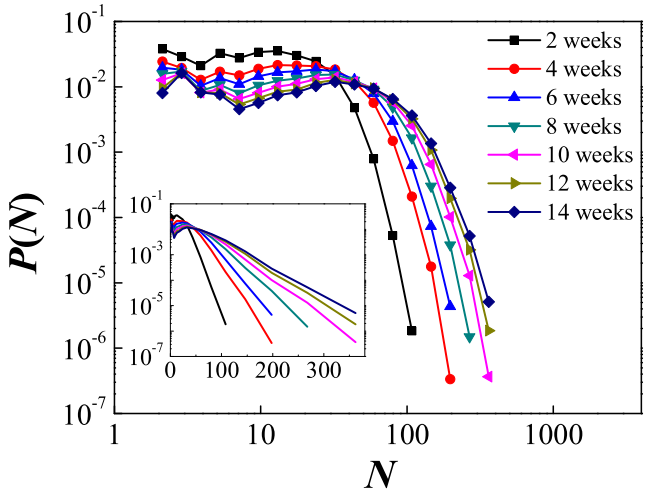


Fig. S1: The distribution of number of locations N for various time periods.

Fig. S1 shows the distribution of the number locations N visited for various windows of time. After three months $P(N)$ converges and can be regarded as relatively saturated, indicating that in this time frame we can discover most of the locations typically frequented by our users. This saturation also indicates that with a good approximation N_i is an accurate estimate of the total number of locations a user visits.

B. Radius of gyration

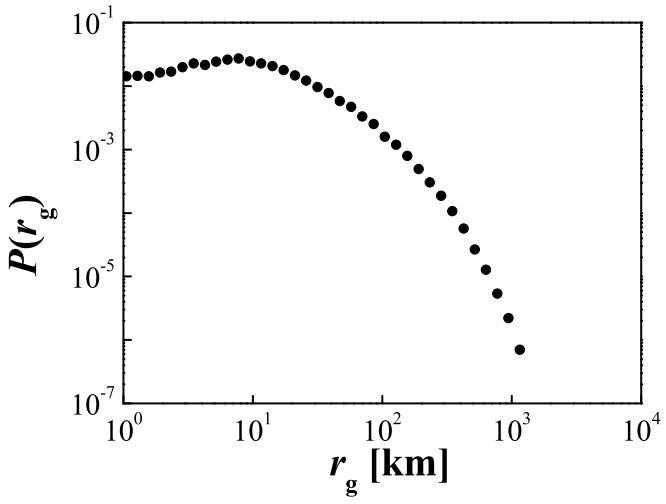


Fig. S2: The distribution of the typical distance covered by each of the $50,000$ users in D_2.

The radius of gyration r_g describes the typical range of a user's trajectory:

$$r_g = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(\vec{r}_i - \vec{r}_{cm})^2}, \tag{S1}$$

where $\vec{r}_i$ represents the position at time i, $\vec{r}_{cm} = \frac{1}{L}\sum_{i=1}^{L}\vec{r}_i$ is the center of mass of the trajectory, and L is the total number of recorded points for the user's location. Fig. S2 shows a fat tailed distribution of r_g for the users considered in this work, reproducing consistent with the results of Ref. [1].

S3. DATA PREPROCESSING

To construct a time series for each user we segment the three month observation period into hour-long intervals. Each interval is assigned a tower ID if one is known (i.e. the phone was used in that time interval). If multiple calls were made in a given interval, we choose one of them at random. Finally if no call is made in a given interval, we assign it an ID "?", implying an unknown location. Thus for each user i we obtain a string of length $L = 24 \times 7 \times 14 = 2352$ with $N_i + 1$ distinct symbols, each denoting one of the N_i towers visited by the user and one for the missing location "?".

S4. DETERMINATION OF USER ENTROPY

In general lower entropy implies higher predictability. Here we discuss how to measure the entropy S of individual mobile phone users over their past history, allowing us to quantify their predictability.

A. Entropy rate and basic equations

Let X_i be a random variable representing a user's location at time i. Entropy is defined as $S = -\sum_{x \in X} p(x) \log_2 p(x)$ where $p(x) = Pr\{X = x\}$ is the probability that $X = x$. The conditional entropy of random variable Y given X is defined as $S(Y|X) \equiv \sum_{x \in X} p(x) S(Y|X = x)$. Let h_n be a random variable for a sequence of n locations. For a stationary stochastic process $\mathcal{X} = \{X_i\}$, the entropy rate may be written as,

$$S \equiv \lim_{n \to \infty} \frac{1}{n} S(X_1, X_2, \ldots, X_n) \tag{S2}$$

$$= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} S(X_i|h_{i-1}), \tag{S3}$$

$$= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} S(i), \tag{S4}$$

where Eq. S2 is the definition of entropy [2], Eq. S3 is an application of the chain rule for entropy, and we define $S(n) \equiv S(X_n|h_{n-1})$ as the conditional entropy at the n-th step in Eq. S4. If the time series lacks any long range temporal correlations (i.e. the probability of the next location is independent of the current one) we have $S = -\sum_i p_i \log_2 p_i$, where p_i

the probability of being at location i.

For an individual visiting N locations, we are interested in the following quantities:

- S_i: the user i's true entropy, considering both spatial and temporal patterns.
- $S^{\text{unc}} = -\sum_{i=1}^{N} p_i \log_2 p_i$: is the temporal-uncorrelated entropy, where p_i is the probability that location i is visited by the user.
- $S^{\text{rand}} = \log_2 N$ is the random entropy, obtained when $p_i = \frac{1}{N}$ for all locations i. In this case each of the N locations is equally probable.

Clearly, $0 \leq S \leq S^{\text{unc}} \leq S^{\text{rand}} < \infty$.

B. Algorithm

To calculate the entropy from the user's past location history, we use an estimator based on Lempel-Ziv data compression [3], which is known to rapidly converge to the real entropy of a time series. For a time series with n steps, the entropy is estimated by

$$S^{\text{est}} = \left(\frac{1}{n} \sum_i \Lambda_i \right)^{-1} \ln n, \tag{S5}$$

where Λ_i is the length of the shortest substring starting at position i which doesn't previously appear from position 1 to $i-1$. It has been proven that S^{est} converges to the actual entropy when n approaches infinity [3].

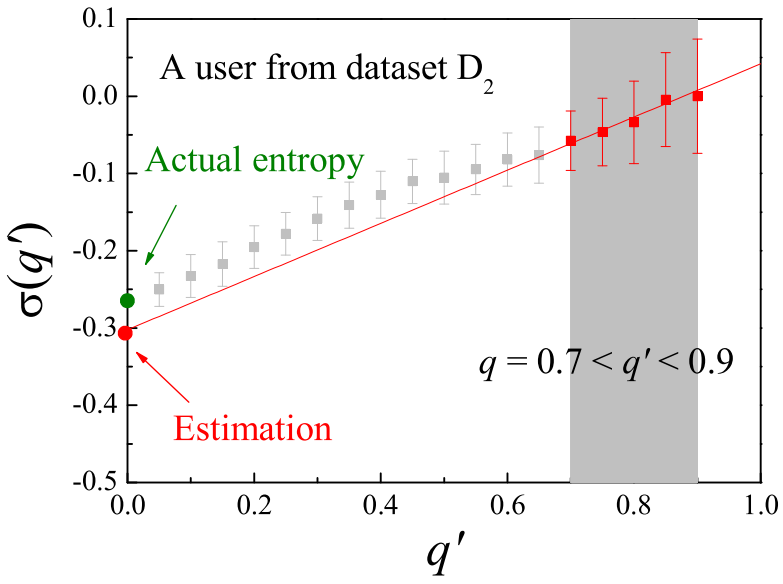


Fig. S3: The order parameter $\sigma(q') \equiv \ln(S(q')/S^{\rm unc}(q'))$ as a function of q' with given $q = 0.7$.

Applying Eq. (S5) to the empirical time series of a user's location history, the obtained entropy $S_i(q)$ will depend on the fraction of unknown locations q. The unknown locations serve as a source of additional entropy $S_i(q)/S_i^{\rm unc}(q) > S_i/S_i^{\rm unc}$, where S_i is the user's true entropy given a complete record of his hourly locations. To determine the true entropy $S_i = S_i(q = 0)$ we use the following algorithm: for a time series with a q fraction of unknown locations we select an additional Δq fraction of known locations and designate them as unknown. That is, we replace a known fraction Δq of locations with ID "?", increasing q to $q' = q + \Delta q$. We then vary $\Delta q = 0, 0.05, 0.10, 0.15, \ldots, 0.90 - q$, measuring the order parameter $\sigma(q') \equiv \ln(s^{\rm eff}(q')) = \ln(S(q')/S^{\rm unc}(q'))$, where $S(q')$ is determined using the Lempel-Ziv algorithm and $S^{\rm unc}(q')$ is the entropy provided by the Lempel-Ziv algorithm over the randomly shuffled time series.

In Fig. S3 we plot $\sigma(q')$ for a typical user from D_2 with $q = 0.7$, observing a reasonably linear relationship between $\sigma(q')$ and q'. Since we cannot directly measure the unbiased case (when $q' = 0$), we extrapolate $S(q')$ from the range $q \le q' \le 0.9$ to $q' = 0$, to estimate σ_{est} at $q = 0$. The entropy is then calculated as $S^{\rm est} = e^{\sigma_{est}} S^{\rm unc}$.

D. Validity of algorithm

To test the validity of our algorithm, we applied this technique to the complete dataset D_2, i.e. to the users whose location history is recorded every hour, thus there is no ambiguity about their hourly whereabouts ($q = 0$). For each user i, we measured the real entropy $S_i^{\rm real}$ using the Lempel-Ziv algorithm. Then we randomly designated q fraction of known locations as "?", artificially mimicking the situation when our dataset is incomplete. Finally we applied the algorithm on the artificially incomplete data, estimating the real entropy $S_i^{\rm est}(q)$. Fig S4 demonstrates how $S^{\rm est}/S^{\rm real}$ varies with the incompleteness fraction q for two typical users in D_2, indicating that the procedure works reasonably well up to $q_\epsilon = 0.7$. As we state below, q_ϵ scales with the length of the time series, thus q_ϵ for the three month dataset D_1 will eventually be larger than the $q_\epsilon = 0.7$ value determined here for the 8-day data.

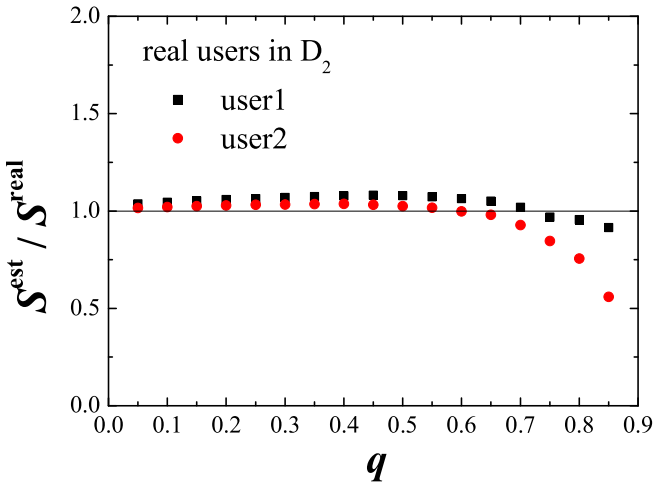


Fig. S4: $S^{\rm est}/S^{\rm real}$ vs q for two different users in D_2.

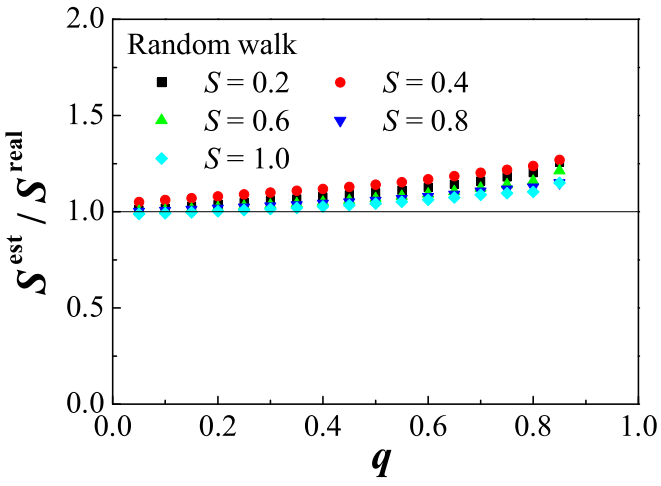


Fig. S5: $S^{\text{est}}/S^{\text{real}}$ vs q for the random model with different values of entropy S.

We also tested our algorithm on a simple two-state random time series. In this case user i visits only two locations (thus $S^{\text{rand}} = 1$). At each time step he visits location 0 with probability p_0 or location 1 with probability $1 - p_0$, thus $S_i = S_i^{\text{unc}} = -p_0 \log_2 p_0 - (1 - p_0) \log_2(1 - p_0)$. In Fig. S5 we plot $S^{\text{est}}/S^{\text{real}}$ vs q for the random model with entropy $S = 0.2, 0.4, 0.6, 0.8, 1.0$ and length $L = 8$ days. As q increases or S decreases, the estimate tends to deviate from the real value, yet the error is less than 25% even for q close to 0.9. In the $q = 0.7$ range, where most of our users are, (see Fig. 2 in the main paper), the error is typically under 10%.

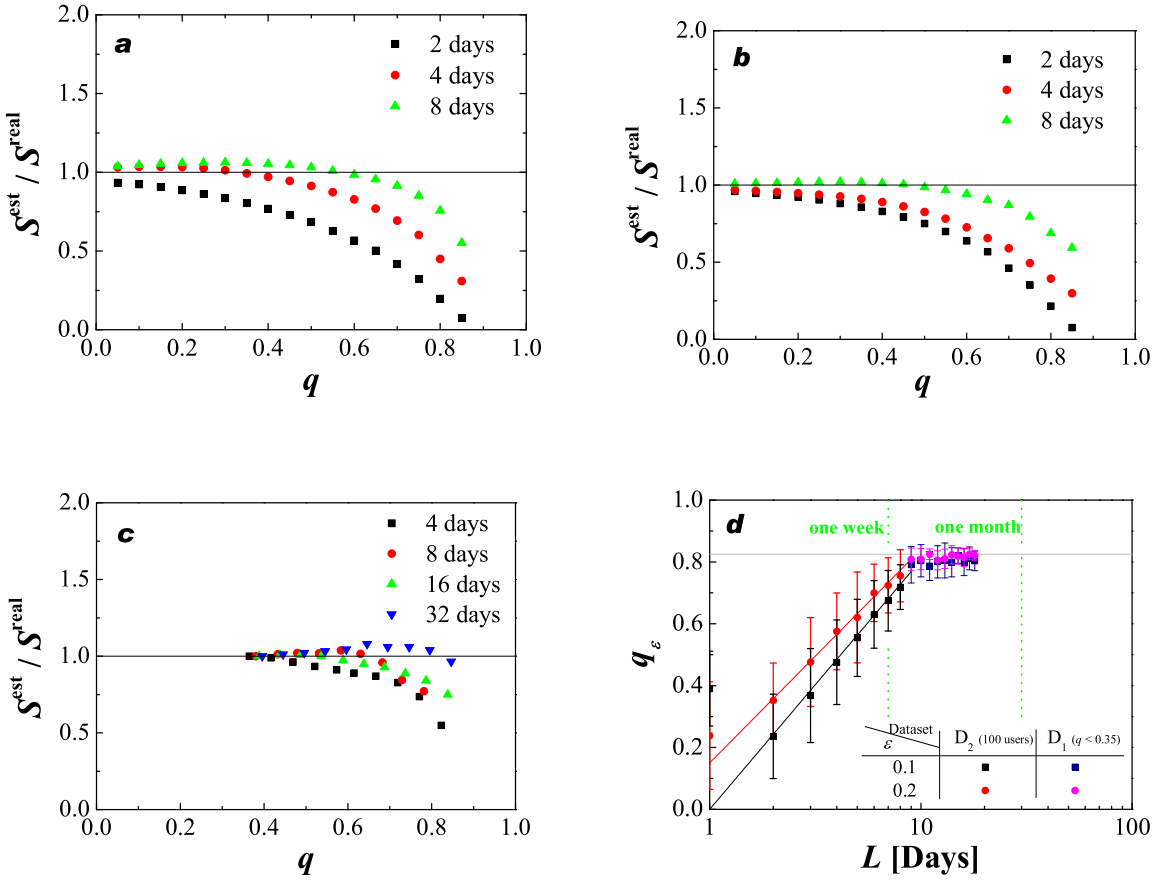


Fig. S6: a, b) $S^{\text{est}}/S^{\text{real}}$ vs q for times series of lengths 2, 4 and 8 days for two typical users in D_2 c) $S^{\text{est}}/S^{\text{real}}$ vs q for times series: 4, 8, 16 and 32 days for a typical user in D_1 with $q = 0.3$. d) The critical q_ϵ defined from the error $|S^{\text{est}}/S^{\text{real}} - 1|$ vs the length of time series for both D_1 (filtered with $q < 0.35$) and D_2, indicating a quick convergence of q_ϵ for long enough time series. The strait line is a logarithmic fitting. The grey line is the threshold of $q_\epsilon^\infty = 0.825$.

It is important to test the validity of our algorithm for different lengths L of the time series. Figures S6a and S6b indicate that the threshold for q increases with L for two users chosen from dataset D_2. Furthermore, we applied the algorithm for users in D_1 with only a small fraction of missing data ($q < 0.35$), thus the entropy measured by the Lempel-Ziv algorithm is roughly equivalent to S^{real}. By increasing the fraction q of missing information, we tested our algorithm up to $q = 0.825$, as shown in Fig. S6c.

To quantify the finite size scaling of the critical value of q, we explicitly define q_ϵ^L as the largest q satisfying $|S^{\text{est}}(q)/S^{\text{real}}(q) - 1| < \epsilon$, where ϵ is the error of the estimation. One may think of q_ϵ^L as a limit to how bad input data of length L can be while still achieving

a good estimate for S^{real}. The upper bound of q_ϵ^L as $L \to \infty$ is limited by the algorithm. Since $q' < 0.9$ and the interval $\Delta q = 0.05$, the maximum possible q' within the fitting region is between 0.85 and 0.9, and thus is 0.875 on average. The linear fitting requires at least two points, which leads to $q_\epsilon^\infty = 0.875 - \Delta q = 0.825$ which represents an upper limit for the algorithm's utility. We then demonstrate the relationship between q_ϵ^L and L for $\epsilon = 0.1$ and 0.2 in Fig. S6d. For $L > 8$ days, we used D_1 with $q < 0.35$ to estimate the q_ϵ^L. We find q_ϵ^L scales with size L logarithmically for small value of L and then converges to $q_\epsilon^\infty = 0.825$ after 20 days. Therefore, for users with $q < q_\epsilon^\infty$ we can determine the entropy with sufficient accuracy. In the following study, we focus on the 45,000 users with $q < 0.8 < q_\epsilon^\infty$, which ensures real entropy S_i for each user i can be accurately determined. Results are presented in Figs. 2 , 3 in the main manuscript.

S5. FUNDAMENTAL LIMITS OF PREDICTABILITY

If a user has entropy $S = 0$, then his/her mobility is completely regular and thus the user's whereabouts is fully predictable. If, however, a user's entropy $S = S^{\text{rand}} = \log_2 N$, then his/her trajectory is expected to follow a random pattern and thus we cannot forecast it with accuracy that exceeds $1/N$. Most users have a finite entropy laying between 0 and S^{rand} however, indicating not only that a certain amount of randomness governs their future whereabouts, but also that there is some regularity in their movement that can be exploited for predictive purposes. In this section we aim to quantify the limits of predictability of a user's next location based on his trajectory history. That is, we want to answer the question: How predictable is a user's next location given the entropy of his trajectory? We will use a version of Fano's inequality to relate the upper bound of predictability to the entropy of a user's past history of mobility. We will also show that the regularity R measured in the main manuscript offers a lower bound to the user's predictability.

A. Notation

Let $h_{n-1} = \{X_{n-1}, X_{n-2}, \ldots, X_1\}$ denote a user's past history from time interval 1 to $n-1$, where X_i corresponds to the user's location at time step i. Let $Pr[X_n = \hat{X}_n | h_{n-1}]$ be the probability that our guess $\hat{X}_n$ for a user's next location agrees with his actual next location X_n given his location history h_{n-1}. Let $\pi(h_{n-1})$ be the probability the user will be

in his most likely next location x_{ML} given his history h_{n-1}. Thus

$$\pi(h_{n-1}) = \sup_x \{ Pr[X_n = x|h_{n-1}] \}, \tag{S6}$$

where $Pr[X_n = x|h_{n-1}]$ is the probability that the next location X_n is x given the history h_{n-1}. That is, $\pi(h_{n-1})$ contains the full predictive power including the potential long-range correlations present in the data.

Let $P_a(\hat{X}_n|h_{n-1})$ be the distribution generated by an arbitrary predictive algorithm a over the next possible location $\hat{X}_n$. Let $P(X_n|h_{n-1})$ be the true distribution over which the user will select his next location. Thus the probability of correctly forecasting the user's next location is $Pr_a\{X_n = \hat{X}_n|h_{n-1}\} = \sum_x P(x|h_{n-1})P_a(x|h_{n-1})$. Since $\pi(h_{n-1}) \geq P(x|h_{n-1})$ for any x, we have

$$\begin{aligned} Pr_a\{X_n = \hat{X}_n|h_{n-1}\} &= \sum_x P(x|h_{n-1})P_a(x|h_{n-1}) \\ &\leq \sum_x \pi(h_{n-1})P_a(x|h_{n-1}) \\ &= \pi(h_{n-1}). \end{aligned} \tag{S7}$$

In other words, any forecasting based on history h_{n-1} cannot do better than the one that places the user in his/her most likely location.

We still must demonstrate that Eq. S7 can in principle be reached, i.e. it represents a tight upper bound. We will show that this maximal predictability is theoretically achievable using a hypothetical algorithm $a^\star$ that has the property

$$P_{a^\star}(x|h_{n-1}) = \begin{cases} 1 & x = x_{ML} \\ 0 & x \neq x_{ML}, \end{cases} \tag{S8}$$

namely $a^\star$ always chooses the user's next most likely location as its prediction. Then

$$\begin{aligned} Pr_{a^\star}\{X_n = \hat{X}_n|h_{n-1}\} &= \sum_x P(x|h_{n-1})P_{a^\star}(x|h_{n-1}) \\ &= \pi(h_{n-1}). \end{aligned}$$

Therefore $\pi(h_{n-1})$ is not only an upper limit, but is in principle attainable by an appropriate algorithm.

Next we define the predictability $\Pi(n)$ for a trajectory that corresponds to a given history of length $n-1$. Let $P(h_{n-1})$ be the probability of observing a particular history h_{n-1}. Then predictability is given by

$$\Pi(n) \equiv \sum_{h_{n-1}} P(h_{n-1})\pi(h_{n-1}), \tag{S9}$$

where the sum is taken over all possible histories of length $n-1$. Taking the limit, we define the overall predictability Π as

$$\Pi \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i}^{n} \Pi(i). \tag{S10}$$

Since $\Pi(n)$ is the best success rate to predict user's location at the time n, Π may be viewed as the time averaged predictability. Next we explore its range.

B. Fano's inequality

Given the $P(X_n|h_{n-1})$ distribution over a user's next location we will create a new distribution that is as random as possible while preserving $\pi(h_{n-1}) = p$ in Eq. (S6). Let N be the total number of possible locations. Keeping p for location x_{ML}, we assume a uniform distribution over the remaining $N-1$ locations. Thus we have X' with an associated distribution $P'(X|h) \equiv \left(p, \frac{1-p}{N-1}, \frac{1-p}{N-1}, \ldots, \frac{1-p}{N-1}\right)$. This distribution is at least as random as the original, thus $S(X_n|h_{n-1}) \leq S(X'|h_{n-1})$. This entropy may be calculated directly as

$$\begin{aligned} S(X'|h_{n-1}) &= -p\log_2(p) - \sum \frac{1-p}{N-1}\log_2\left(\frac{1-p}{N-1}\right) && \text{(S11)}\\ &= -p\log_2(p) - (1-p)\log_2\left(\frac{1-p}{N-1}\right) && \text{(S12)}\\ &= -\left[p\log_2 p + (1-p)\log_2(1-p)\right] + (1-p)\log_2(N-1) && \text{(S13)}\\ &\equiv S_F(p) = S_F(\pi(h_{n-1})). && \text{(S14)} \end{aligned}$$

Note that

$$S(X_n|h_{n-1}) \leq S_F(\pi(h_{n-1})), \tag{S15}$$

which represents an appropriate rewriting of Fano's inequality [2].

It is important to realize that for $p \in [1/N, 1)$ the Fano function $S_F(p)$ is concave and monotonically decreases with p. That is, $S_F((a+b)/2) \geq (S_F(a)+S_F(b))/2$ (*concavity*) and

$(S_F(a) - S_F(b))(a - b) \leq 0$ (*monotonically decreasing*), as shown in Fig. S7.

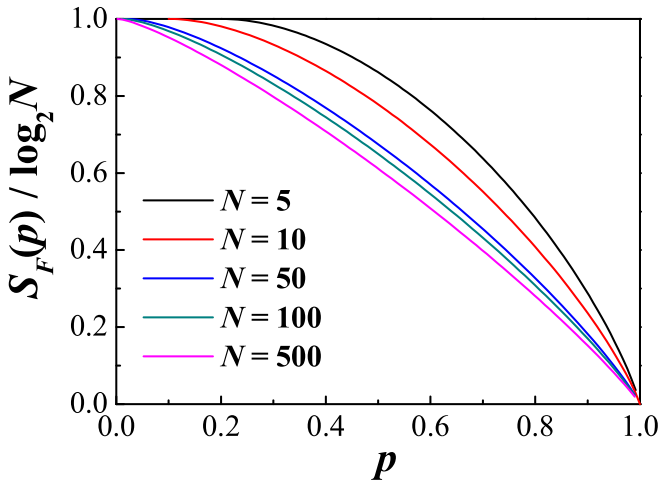


Fig. S7: The plot of Fano function S_F normalized by the maximum entropy $S^{\text{rand}} \equiv \log_2 N$, showing that S_F is a concave and monotone decreasing function.

C. Upper bound of predictability $\Pi^{\max}$

We wish to relate the entropy rate S defined in Eq. S4 which we estimated using the algorithm developed in Section S4 to predictability Π defined in Eq. S10.

We begin with the simpler case of relating the conditional entropy $S(n) = S(X_n|h_{n-1})$ to $\Pi(n)$ as defined in Eq. S9. Recall that h_{n-1} represents a history of length $n-1$ and $P(h_{n-1})$ is the probability of observing the particular history h_{n-1}. Then

$$S(n) = \sum_{h_{n-1}} P(h_{n-1}) S(X_n|h_{n-1}) \tag{S16}$$

$$\leq \sum_{h_{n-1}} P(h_{n-1}) S_F\left(\pi(h_{n-1})\right) \tag{S17}$$

$$\leq S_F\left(\sum_{h_{n-1}} P(h_{n-1})\pi(h_{n-1})\right) \tag{S18}$$

$$= S_F\left(\Pi(n)\right) \tag{S19}$$

Here Eq. S16 is the definition of conditional entropy. Eq. S17 follows from Eq. S15. Eq. S18 follows from Jensen's inequality and the fact that S_F is concave in p. Eq. S19 follows from our definition of $\Pi(n)$.

We now have $S(n) \leq S_F(\Pi(n))$ to which we will again apply Jensen's inequality to obtain a relationship between S and Π.

$$S = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} S(i) \tag{S20}$$

$$\leq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} S_F(\Pi(i)) \tag{S21}$$

$$\leq S_F \left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Pi(i) \right) \tag{S22}$$

$$= S_F(\Pi). \tag{S23}$$

Here Eq. S20 is from the definition of entropy in Eq. S4. Eq. S21 follows from Eq. S19. Eq. S22 follows from Jensen's inequality and that S_F is concave. Eq. S23 follows from the definition of Π in Eq. S10.

Now let's define $\Pi^{\max}(S, N)$ as the solution to the equation

$$\begin{aligned} S &= S_F(\Pi^{\max}) \\ &= -\left[\Pi^{\max} \log_2 \Pi^{\max} + (1-\Pi^{\max}) \log_2 (1-\Pi^{\max})\right] + (1-\Pi^{\max}) \log_2 (N-1) \\ &\leq S_F(\Pi) \end{aligned} \tag{S24}$$

where Eq. S24 follows from Eq. S23.

Based on the fact that $S_F(\Pi^{\max}) \leq S_F(\Pi)$ and $S_F(\Pi)$ monotonically deceases with Π, we have

$$\begin{aligned} \left[S_F(\Pi^{\max}) - S_F(\Pi)\right](\Pi^{\max} - \Pi) &\leq 0 \\ \Pi^{\max} - \Pi &\geq 0 \\ \Pi^{\max} &\geq \Pi. \end{aligned}$$

In other words $\Pi^{\max}$ represents an upper bound of predictability Π.

D. Regularity as a lower bound of predictability

As we try to establish a lower bound for the user's predictability, we consider the most likely location x'_{ML} at a specific time of day. Thus rather than considering the entire history and the potential correlation in the mobility pattern, we only look at where the user was for

example on Monday between 9AM and 10AM. There exists a *set* of possible true histories that will be consistent with our observed behavior for Monday morning.

Imagine we know a user's location every Monday at 10AM. We will call this string of locations $C = x_1, x_2, \ldots$. There exists many possible histories that can satisfy constraint C. For example if x_1 is the office, there are many possible trajectories one can take to get to the office, as long as he is there by 10AM Monday. Let h'_{n-1} be an element in the set of all such histories satisfying constraint C.

We define $R(n)$, or *regularity* at the n-th step as the expected $\pi(h'_{n-1}) \equiv P(x'_{ML}|h'_{n-1})$ over all constrained histories h'_{n-1}. Next we will show that $R(n)$ represents a lower bound for $\Pi(n)$. Each of the following steps is explained below.

$$\Pi(n) \equiv \sum_{h_{n-1}} P(h_{n-1})\pi(h_{n-1}) \tag{S25}$$

$$= \sum_{h_{n-1}} \left(\sum_{h'_{n-1} \in H^C} P(h'_{n-1}) P(h_{n-1}|h'_{n-1}) \right) \pi(h_{n-1}) \tag{S26}$$

$$= \sum_{h'_{n-1} \in H^C} P(h'_{n-1}) \left(\sum_{h_{n-1}} P(h_{n-1}|h'_{n-1}) \pi(h_{n-1}) \right) \tag{S27}$$

$$\geq \sum_{h'_{n-1} \in H^C} P(h_{n-1}) \pi(h'_{n-1}) \tag{S28}$$

$$= R(n). \tag{S29}$$

Eq. S25 is the definition of $\Pi(n)$. Eq. S26 is based on the identity $\sum_{h'_{n-1} \in H^C} P(h'_{n-1}) P(h_{n-1}|h'_{n-1}) = P(h_{n-1})$. Eq. S27 is exchanging the summing over h_{n-1} and h'_{n-1}. Eq. S28 is because for any location x we have

$$P(x|h'_{n-1}) = \sum_{h_{n-1}} P(h_{n-1}|h'_{n-1}) P(x|h_{n-1}) \leq \sum_{h_{n-1}} P(h_{n-1}|h'_{n-1}) \pi(h_{n-1}), \tag{S30}$$

thus for most likely location $x = x'_{ML}$ and $\pi(h'_{n-1}) = P(x = x'_{ML}|h'_{n-1})$. Eq. S29 is our definition of $R(n)$.

Now we define the time averaged regularity as

$$R \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} R(i) \leq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Pi(i) = \Pi \tag{S31}$$

Combining this result with the upper bound, the predictability Π of a user satisfies $R \leq \Pi \leq \Pi^{\max}$. Note, however, that R represents a rather generous lower bound as it ignores potential long range correlations in the user's travel patterns, which could have considerable predictive power.

S6. REGULARITY ON WEEKDAYS AND WEEKENDS

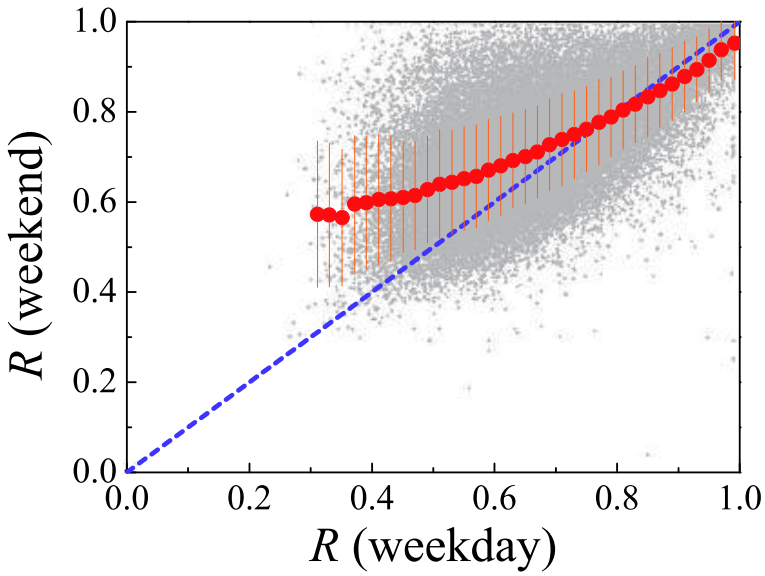


Fig. S8: Regularity on weekdays vs. weekends across the user base. The gray dots correspond to each of the 45,000 users. The red symbols are the averaged trend.

Due to the lack of work related constrains people are expected to show a higher degree of spontaneity and thus are less predicability over the weekends. To test this hypothesis, in Fig. S8 we measure the regularity for each individual during weekdays and weekends, respectively. Surprisingly, we do not find significant changes in the user's mobility pattern over the weekend. To the contrast, 65% of users exhibit greater regularity during the weekend than during the weekdays (the data points above the blue dashed line). The average trend shows (red symbols) that only the users with very high regularity ($R > 0.8$) have a decreased average regularity during the weekend. Note that only 19% of the users have $R > 0.8$.

This suggests that it is not the regularity imposed on us by the work schedule that keeps us predictable, rather we are potentially capturing something intrinsic to human activity, that spans both weekdays and weekends. People who have a desire for regularity tend to exhibit that throughout the weekday and weekend, perhaps making both professional and recreational choices accordingly.

S7. THE DEMOGRAPHIC DEPENDENCE

A. Dependence on number of locations distinct N visited by the user

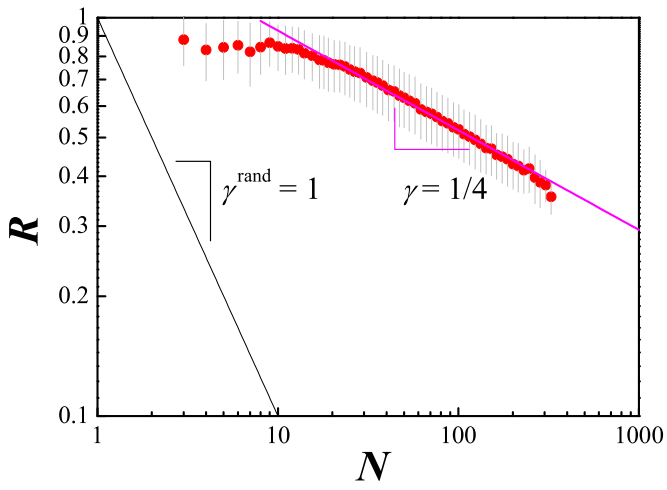


Fig. S9: The regularity R vs the number of locations N, showing that R decays slowly with N and $R(N) \sim N^{-1/4}$.

Fig. S9 shows R deceases with N as $R(N) \sim N^{-1/4}$. This is a much slower decay than the random case obtained if we assume that each of the N locations has equal probability and thus $R^{\text{rand}}(N) \sim N^{-1}$.

B. Age and gender dependency

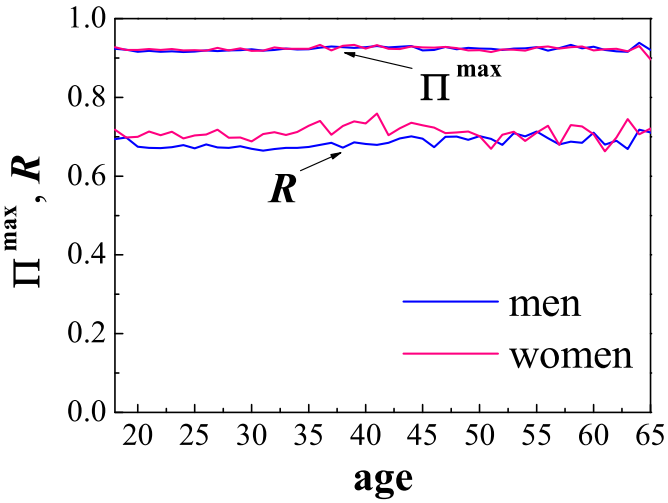


Fig. S10: The dependence of the maximal predictability $\Pi^{\max}$ and regularity R on the age of the users, shown separately for men (blue) and women (red).

Figure S10 indicates that no any gender or age based differences on the potential predictability $\Pi^{\max}$ whereas women have slightly higher regularity than men.

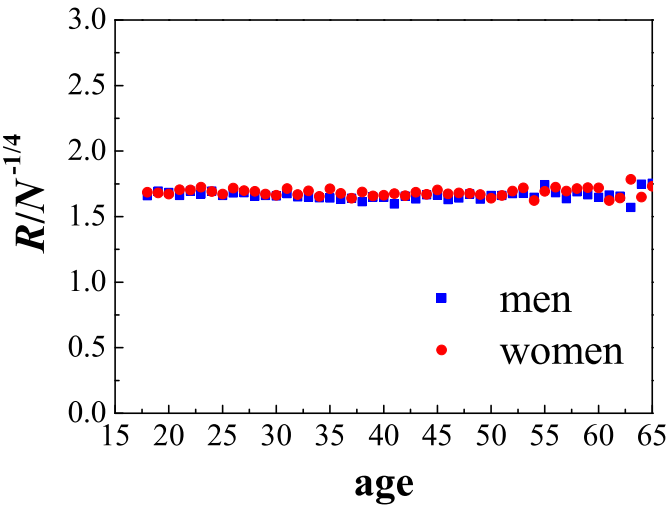


Fig. S11: The dependence of the normalized regularity $R/N^{-1/4}$ on the age of the users, shown separately for men (blue) and women (red).

This gender dependency is rooted in the N-dependency of regularity. Indeed, if we normalize the regularity R by $N^{-1/4}$ obtained in the previous section (Fig. S9), the gender dependency vanishes, as shown in Fig. S11.

C. Dependence on the income and language

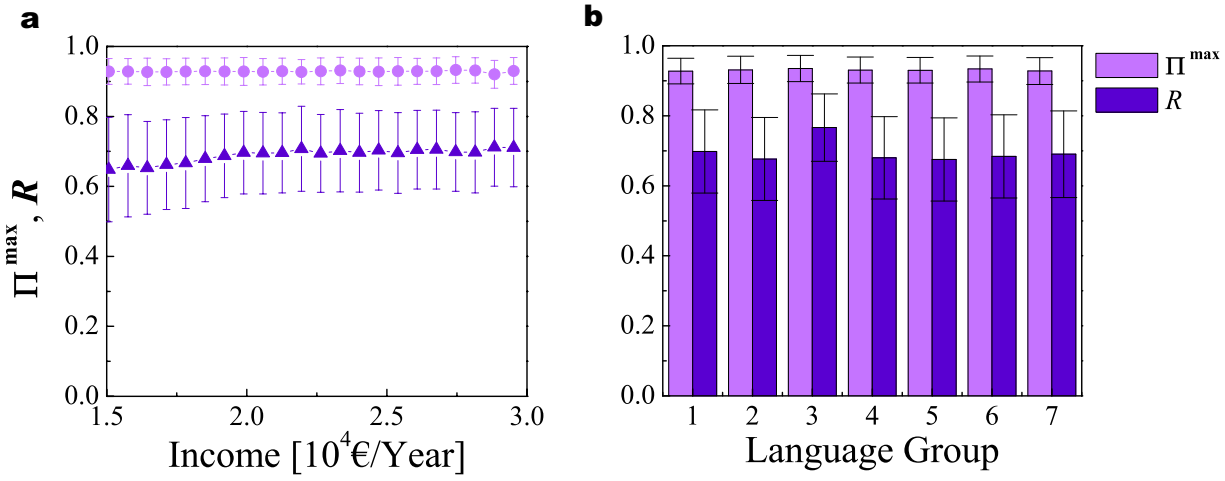


Fig. S12: Predictability is stable across all regional income levels and language. Users were assigned a province based on their most-used tower. (a) Provinces were assigned a mean annual income based on census data. (b) Provinces were assigned a regional language if one exists, otherwise they were assigned the national language.

Using census data, we are able to assign average income and language to the users based on their most visited location. As Fig. S12 shows, while the regularity R (the lower bound of predictability) appears to depend somewhat on the various demographic parameters, the maximum predictability $\Pi^{\max}$ does not, showing only small fluctuations. Note that ideally we should assign these parameters to individual users, thus more definite answer could be possible once such microscopic (user-specific) data becomes available.

D. Dependence on the population density

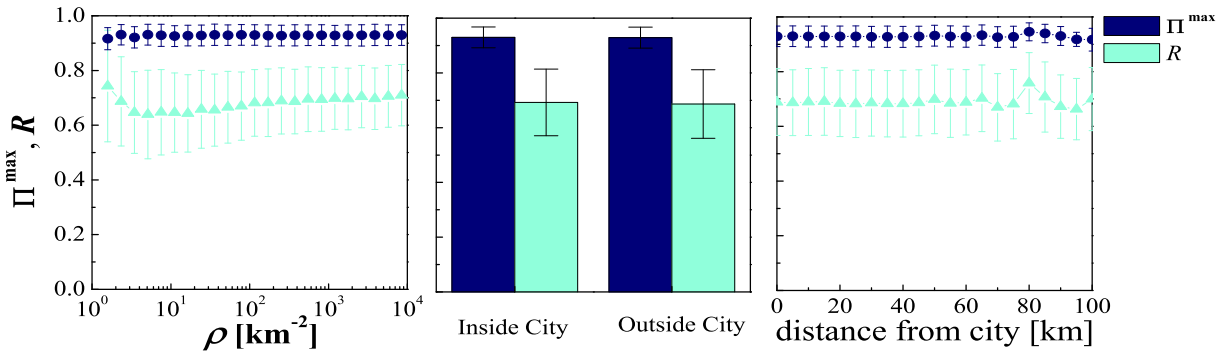


Fig. S13: (a) The dependence of the maximal predictability $\Pi^{\max}$ and regularity R on the population density ρ (the number of people per km^2) of the users' neighborhood (11,177 neighborhoods based on the zip code). (b) The predictability $\Pi^{\max}$ and regularity R inside or outside metropolises, which were the four most populated cities in the country. (c) The predictability $\Pi^{\max}$ and regularity R vs. the distances from the top four cities.

It is important to explore if predictability depends on population density. For this we have identified for each user his/her most frequented location, and using census data we assigned to the user a population density specific to the region that the user most frequently visits. Fig. S13 shows that despite the changes in the population density that spans four orders of magnitude, user predictability is largely constant. We observe small changes only in the regularity R.

---

[1] Gonzalez, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779-782 (2008).

[2] Cover, T. M., Thomas, J. A. *Elements of Information Theory* (John Wiley & Sons, Hoboken, NJ, 2006).

[3] Kontoyiannis I., Algoet P. H., Suhov Yu. M., Wyner A. J. Nonparametric Entropy Estimation for Stationary Processes and Random Fields, with Applications to English Text, *IEEE Transactions on Information Theory* **44**, 1319-1327 (1998).

[4] Navet N., Chen S-H. On Predictability and Profitability: Would GP Induced Trading Rules be Sensitive to the Observed Entropy of Time Series? *Natural Computing in Computational Finance* **100** 197-210 (2008).