Zhao et al. *EPJ Data Science* (2021) 10:5
https://doi.org/10.1140/epjds/s13688-021-00261-2

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Characteristics of human mobility patterns revealed by high-frequency cell-phone position data

Chen Zhao[1,2,3], An Zeng[4*] and Chi Ho Yeung[5]

*Correspondence:
anzeng@bnu.edu.cn
[4]School of Systems Science, Beijing Normal University, 100875 Beijing, P.R. China
Full list of author information is available at the end of the article

Abstract

Human mobility is an important characteristic of human behavior, but since tracking personalized position to high temporal and spatial resolution is difficult, most studies on human mobility patterns rely on sparsely sampled position data. In this work, we re-examined human mobility patterns via comprehensive cell-phone position data recorded at a high frequency up to every second. We constructed human mobility networks and found that individuals exhibit origin-dependent, path-preferential patterns in their short time-scale mobility. These behaviors are prominent when the temporal resolution of the data is high, and are thus overlooked in most previous studies. Incorporating measured quantities from our high frequency data into conventional human mobility models shows inconsistent statistical results. We finally revealed that the individual preferential transition mechanism characterized by the first-order Markov process can quantitatively reproduce the observed travel patterns at both individual and population levels at all relevant time-scales.

Keywords: Human mobility; Mobile phone; High frequency data

1 Introduction

Due to the increasing availability of mobile-phone records, global-positioning-system data and other datasets capturing traces of human movements, numerous statistical patterns in human mobility have been revealed, ranging from the confined radius of gyration at the individual level [1] to the commuting fluxes at the collective level [2]. These empirical observations suggest that human mobility are barely random, but follow predictable rules [3–12]. Accordingly, models have been proposed to understand the observed mobility patterns. Following the pioneer model which generates empirical scaling behaviors by introducing two generic mechanisms, exploration and preferential return (EPR) [2], a large number of models for individual human mobility have been developed. Examples include the variants of the EPR model which describe user virtual mobility in cyberspace [13–15] by incorporating a gravity model to simulate the returner-explorer dichotomy [16], introducing a social circle to model the conserved number of locations an individual visits [17], aggregating individual trajectories to generate collective movements [18], and so on.

Springer

On the other hand, it has been shown that there is a diversity of human mobility patterns at different spatial scales. On the largest spatial scale which constitutes international movements, they are largely constrained by the entry requirement of individual countries, leading to asymmetric international movements [19, 20]. On the spatial scale within a country, models which describe international movements do not well explain inter-city movements. For instance, the inter-city human mobility is claimed to be mainly driven by the search for better job opportunities [6, 21]. The radiation model assumes that individuals tend to select the nearest locations with large benefits. On the spatial scale within a city, the local movements are better predicted by a population-weighted opportunity model where the potential area of coverage of individuals includes the whole city as a manifestation of the high mobility at the city scale [22]. Although large efforts have been devoted to understand human mobility at different spatial scales, the studies of human mobility at different temporal scales are limited, due to the lack of high frequency mobility data [23, 24]. Understanding spatial-temporal human mobility patterns at different scales would lead to numerous applications, such as suppressing epidemic spreading [25, 26], mitigating traffic congestion [27, 28], urban planning [29, 30] and so on.

To reveal the human mobility pattern at different temporal scales, high frequency position data are required. While most existing empirical studies on human mobility are based on cell phone position data, these data are CDRs (Call Detail Records) where user positions are only recorded when they initiate or receive a call or a text message [31]. These datasets can include position records of up to several million anonymous mobile phone users, but the data has in general a low temporal resolution, as user positions are not recorded most of the time. There is a recent work pointing out that position sampling frequency may significantly alter some statistics of human mobility [32]. The missing position data in some literature are interpolated via specific optimization algorithms or are incorporated from other data sources [33, 34]. Difference may exist between the interpolated and the real data. Another usual practice to improve the temporal resolution of the data is to filter out users with long idle periods. For instance, this approach has been applied to extract a sample of user data with sufficient mobility records for inferring the nature of their visited locations such as home and workplace, and their tour trajectories with start and end point at home are investigated accordingly [28]. However, many problems still remain. On one hand, the user filtering procedure may lead to the risk of biased sampling of the original data. Specifically, the filtered data only include users who make frequent phone calls and may be biased to users with specific professions. On the other hand, the temporal resolution of the data after filtering is still insufficient (as frequent as every 10 min in existing literature), leaving many detailed user mobility traces missing from the data. Another possible data source is global-positioning-system (GPS) data [35, 36]. Their temporal resolution can be very high, but as GPS data are mostly recorded by navigation devices in vehicles, it only records positions when users are driving. As a result, GPS data are commonly used for analyzing traffic [37].

In this paper, we utilize the cell phone 4G communication data in Shijiazhuang, a city in northern China, to identify the location of individual cell phone user to a high frequency of every second. With this high-frequency position data, we study human mobility patterns at different time-scales. We find that human show a low tendency to re-visit locations that one has frequently visited. Instead, individuals exhibit origin-dependent, path-preferential patterns in their short time-scale mobility. Finally, we consider a simple model character-

ized by the first-order Markov process to quantitatively reproduce the observed travel patterns at both the individual and population levels in the high temporal resolution data. Our work reveals the heterogeneity in human mobility mechanism at different temporal scales, opening up a new dimension for understanding human mobility behaviours.

2 Data

Our study is based on a full set of 4G communication data for 14 days between cell phones and cell towers in Shijiazhuang, the capital and largest city of North China's Hebei Province. The city has population over 10 million, and its total area is 15,848 square kilometers (Urban area is 2206 square kilometers). There are about 12,000 4G cell towers in Shijiazhuang, with 7000 towers in urban area and 5000 towers in suburb area. The position of a user is recorded when his/her cell phone connects to the closest cell towers for the 4G communication service [38]. As most applications in cell phones constantly exchange data with the back-end servers, the position of a user can be recorded up to every second. Compared with the traditional cell phone data (CDRs) where the position of users is only recorded when they make phone calls, our obtained dataset is much higher in temporal resolution for analyzing individual mobility behavior.

Due to the popularity of smart phones, our dataset actually covers a large proportion of population in the city. For privacy reasons, the data is anonymous and each user is assigned with a unique ID. The original data include records of 5,336,194 users. In order to obtain a dataset describing the mobility patterns of active users with high temporal resolution, we have implemented strict rules to exclude users who do not move at all and those whose data is largely incomplete (i.e. those who have one or more days with less than 20-hour daily record in the consecutive 14-day period). Finally, we single out and analyze the mobility data of 55,389 users who satisfy the above criteria. The basic descriptive statistics of this data is shown in Fig. S1 and S2 of the supplementary information (SI, see Additional file 1).

3 Results

Empirical human mobility pattern at short time-scales. We start our analysis by constructing the mobility network of a typical mobile phone user in Fig. 1a. Each node is a location defined by an area of the geographical location of the cell tower. The network only consists of the nodes visited and stayed more than 3 minutes by the user, with node size proportional to the frequency he/she visited the location. Two nodes are connected by a link if the user has traveled at least once between the two locations. To understand the mobility patterns in the high temporal resolution data, we shuffle the trajectory of typical users by randomly reordering the sequence of their visited locations. The frequency users visited specific locations is therefore preserved. The mobility pattern constructed from the shuffled trajectory of the typical user in Fig. 1a is illustrated in Fig. 1b. An obvious difference is observed when we compare Fig. 1a and 1b, suggesting that preserving the visitation frequency of locations fails to reproduce mobility networks obtained with the high frequency dataset. Similar results of the real and the shuffled trajectories of three other randomly selected users are shown in Fig. S3 of the SI.

In order to quantify the statistical difference between the mobility patterns in real and shuffled trajectories, we consider four metrics to quantify the trajectories of individuals.

The first one is the total number of unique transited location pairs (transited pairs for short), denoted as n_α^{pair} for user α, which is equivalent to the number of links in the mobility

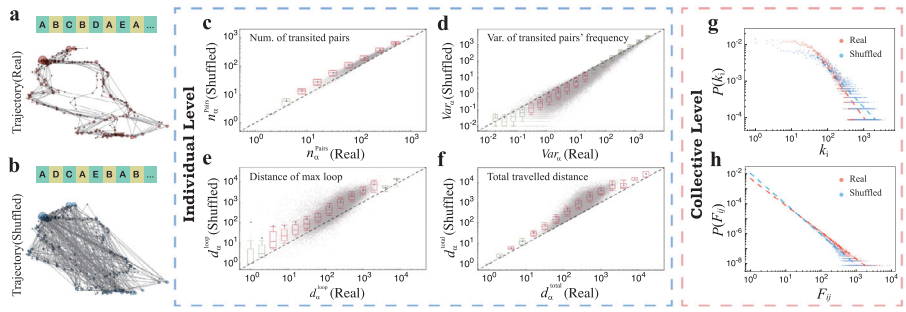

Figure 1 Comparison of the statistical properties of the human mobility in the real and the shuffled data. An illustration to compare (**a**) the real and (**b**) the shuffled travel trajectory of a typical cell phone user. Here each node is a location visited by the user, with node size proportional to its visitation frequency. A link is drawn when the user has traveled at least once between the two locations. The shuffled data is obtained by randomly reordering the sequence of visited locations. In this way, the visitation frequency of each location by the user is preserved while the travel trajectory is randomized. (**c**-**f**) Scatter plots comparing the statistical properties between the real data and the shuffled data for each user α at the individual level, in terms of (**c**) the total transited location pairs n_α^{pairs}, (**d**) the variance Var_α of the traveled frequency of location pairs, (**e**) the covered distance d_α^{loop} of the maximum loop, and (**f**) the total traveled distance d_α^{total}. A box in the standard boxplots are marked in green if the line $y = x$ lies between 10% and 91% in each bin and in red otherwise. (**g**-**h**) Comparison of the statistical properties of the real data and the shuffled data at the collective level, in terms of the distributions of (**g**) the number of neighboring locations of each location, $P(k_i)$ (see (**a**) for example), and (**h**) the distribution of population flux between each two locations, $P(F_{ij})$

network of user α. We then compare n_α^{pair} for all users in the real data and the shuffled data in Fig. 1c. A box in the standard boxplots are marked in green if the line $y = x$ lies between 10% and 91% in each bin and in red otherwise. One can see that n_α^{pair} in the shuffled data is significantly larger than that in the real data. It is because for each individual there exists a few locations with large visitation frequency (e.g. home or office), in the shuffled data users are attracted back to these locations regardless of the distance from the current location, before visiting other locations. In the real data, however, users do not always return to the frequently visited locations if they are too far away, resulting in a much smaller n_α^{pair}, i.e. a much fewer transited pairs than that in the shuffled data.

The second metric we examined is the spread, as measured by the variance Var_α, among the usage frequency of transited pairs of user α (i.e. link weights in the mobility network). As shown in Fig. 1d, a large Var_α indicates that an individual α repeatedly uses a small number of routes and occasionally traveled through other routes. One can see that the values of Var_α are larger in the real data than in the shuffled data, implying that users in the real data more frequently travel between a smaller number of location pairs.

The third metric we examined is the covered distance d_α^{loop} of the maximum loop travelled by user α. Here, a loop is defined as a trajectory that an individual starts from one location and ends in the same location. As shown in Fig. 1e, d_α^{loop} is computed as the total geographic distance of the longest loop in each user's mobility trajectory. Larger d_α^{loop} is observed in the shuffled data, as users in the shuffled data always return to the frequently visited locations even if they are far away.

Finally, the fourth metric, the total traveled distance d_α^{total}, is larger in the shuffled data, as shown in Fig. 1f. As this metric is very sensitive to discrepancies in the predicted trajectory, it is largely ignored in the existing literature. The larger d_α^{total} in the shuffled data is also due to the fact that users often return to the far away yet frequently visited locations in the shuffled data. In fact, d_α^{total} is an important metric, capturing the geographic features

of human mobility. All the above results suggest that although the shuffled trajectories of individuals preserved the location visitation frequency, the patterns from the shuffled data are significantly different from those in the real data.

We further investigate the effect of shuffling on the human mobility patterns at the collective level. From each location, we compute the number of different locations that users travel to. This quantity is essentially the number of links that a location i has in the mobility network, denoted as k_i. The corresponding distribution is shown in Fig. 1g. We see that both distributions $P(k_i)$ of the real and the shuffled data resemble distributions with a power-law tail, yet their exponents are clearly different, with the tail obtained from the real data to be much shorter. The exponents are respectively −1.27 for the real data and −1.05 for the shuffled data, obtained by power-law fitting to the tail of the distributions starting from $k_i = 50$. We see similar difference when we compare the population flux F_{ij} between each pair of locations ij in the real data and the shuffled data in Fig. 1h. Both the flux F_{ij} in the real data and the shuffled data follow power-law distributions. However, the exponent for the fitted power-law function is larger in the real data, indicating that the distribution $P(F_{ij})$ of the real data has a longer tail and a larger maximum value of F_{ij}. The exponents are respectively −1.81 for real data and −1.91 for shuffled data, obtained by power-law fitting to the whole distributions of $P(F_{ij})$. For both $P(k_i)$ and F_{ij}, we have also fitted the probability distributions after log-binning, and obtained the similar exponents as presented above (See Fig. S4 in SI).

Other than revealing human mobility patterns in the spatial dimension, our high frequency data also allow us to reveal the temporal dimension of human mobility activities. To this end, we denote the duration of each of a user's stay at a location as t^{stay}, and examine the distribution $P(t^{\text{stay}})$ over all users. As we can see in Fig. 2a, $P(t^{\text{stay}})$ shows a power-law head and an exponential tail. The power-law function with exponent −1.41 has been used to fit the head of the distribution until $t^{stay} = 6$ (hours). The power-law head suggests that the duration of a stay at different locations is heterogeneous, and there are a large number of locations with relatively short duration of each stay. Note that these values of duration are sufficiently large, e.g. larger than 3 minutes (typical time for users to walk out of the several hundred meters radiation range of a cell tower), and are not pass-by locations. On the other hand, the small peak at the tail is mostly contributed by the duration when users stay or sleep at home.

As evident from Fig. 2a, many locations visited by users for a short time may have been neglected if the dataset do not have a high temporal resolution. Since our 4G cell phone data record user positions in every second, this allows us to examine data with different temporal resolution by data pruning. In order to examine how the mobility statistics are affected by the temporal resolution of the datasets, we consider a threshold and remove all the visited locations with $t^{\text{stay}} < T$, for all users. In Fig. 2b, we show the average number of visited locations as a function of T. One can see that the number of visited locations decreases with an increasing T in a power-law form with an exponent −0.73, implying that the lower the temporal resolution of the data, the more substantial fraction of the visited locations are overlooked in the analyses. Indeed, many hidden mobility patterns at the short time-scale may have been neglected in existing studies which are based on mobility datasets with a low temporal resolution.

To further examine how the temporal resolution of the dataset affects the mobility statistics, we show in Fig. 2c-2f the difference between the real and the shuffled data in terms of

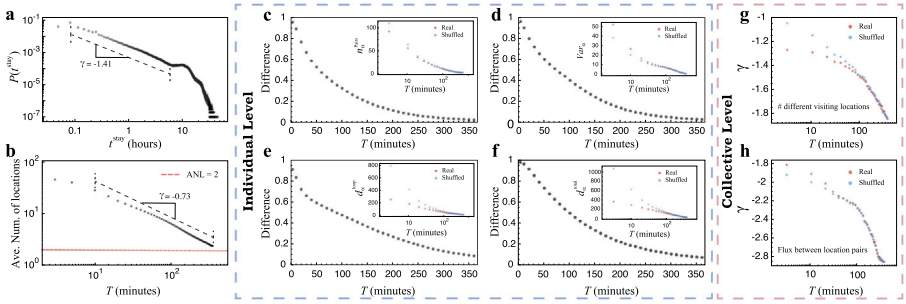

Figure 2 The dependence of mobility statistical properties on the data temporal resolution. (**a**) Distribution $P(t^{\text{stay}})$ of the duration t^{stay} of a stay at each location for all users. We then conduct the statistics below by removing the visited locations with t^{stay} for all users. (**b**) The average number of visited locations as a function of T. In fact, the threshold T can be regarded as a parameter controlling the resolution of the data. A larger threshold T identifies less visited locations, and thus corresponds to a poorer data resolution. (**c**-**f**) The difference between the real data and the shuffled data in terms of (**c**) the total transited location pairs n_α^{pairs}, (**d**) the variance Var_α of the traveled frequency of location pairs, (**e**) the covered distance d_α^{loop} of the maximum loop, (**f**) the total traveled distance d_α^{total}, as a function of threshold T. The insets in each figure show the mean values of the corresponding metric under different threshold. The power-law exponents γ of the distributions of (**g**) the number of neighboring locations from each location, and (**h**) population flux between each two locations, in the real data and the shuffled data as a function of threshold T. In both (**g**) and (**h**), there is large difference between the exponents of the real and the shuffled data when the threshold T is small, and the difference becomes negligible when the threshold is large

the total number of transited location pairs n_α^{pair}, the spread Var_α of the traveled frequency of transited pairs, the distance d_α^{loop} covered by the maximum loop, the total traveled distance d_α^{total}, under various data removal thresholds T. The difference is measured by the fraction of users whose metric values in the shuffled data are larger than those in the real data, except for Var_α. As data shuffling tends to decrease the spread of the traveled frequency of transited pairs, the difference in Var_α is computed as the fraction of user α with Var_α in the shuffled data smaller than that in the real data. Remarkably, when temporal resolution is low (i.e. T is large), our results only show a small difference between the real and the shuffled data in terms of these four metrics at the individual level. As a lower temporal resolution which corresponds to a smaller number of links in the network, it is important to check whether the observed small difference is an artifact of the shuffling process which cannot alter the network structure in these small networks. We show in Fig. S5 in SI that when T is large (e.g. even for T = 150), there are multiple nodes and it is still possible for the shuffling process to change the structure of the network. In addition, we generate a random network with the same number of nodes and same degree sequence as the real network for each threshold T. We find that the difference of the network before and after shuffling is almost constant with respect to T, given the initial network is randomly generated. In comparison, the differences between the real and shuffled network are much higher and are decreasing with respect to T, indicating that the difference between the original and shuffled real networks is not an artifact of increasing T (see Fig. S5 and note 3 in SI).

Similar results can be observed when we compare the power-law distributions in Figs. 1g and 1h under different temporal resolutions. Figures 2g and 2h show that the difference between the exponents of the distributions in Figs. 1g and 1h obtained from the real and the shuffled data is large when the threshold T is small, then become negligible when T is large. Another important observation in Fig. 2h is that the exponent magnitude of the

flux distribution increases with T, indicating that the maximum flux between locations is higher in cases with large threshold. In other words, using datasets with a low temporal resolution would underestimate the flux between locations. Additionally, we study motifs in human travel trajectories [39] in Figs. S6 and S7 (see discussion in SI note 4). A detailed comparison of the human travel motifs in the real data and shuffled data shows that the shuffling process does not significantly alter the motif distribution when T is large, yet the difference between the motif distribution in the real data and the shuffled data is substantial when T is small.

Origin-dependent preference on the next visiting location. In order to understand the reasons underlying the observed difference between the real data and the shuffled cases, we compare their matrices recording the travel frequency of a typical user between each location pair. The matrices are computed with the temporal resolution T = 3 min, and are shown as heatmaps in Figs. 3a and 3b respectively for the real data and the shuffled data. Some large values can be seen in the heatmap of the real data, which suggests that users tend to repeatedly transit between a small number of location pairs. However, this preference of transitions, or equivalently the preference of transited location pairs, cannot be captured in the shuffled data.

We further examine the probability for the selected typical user to visit different locations starting from different origins in Fig. 3c. Different locations are indexed in the horizontal axis, with each blue curve corresponds to the probability to visit other locations from a specific origin; the black dashed curve corresponds to the overall visitation probability distribution. Compared to the black dashed curve, different blue curves peak at dif-

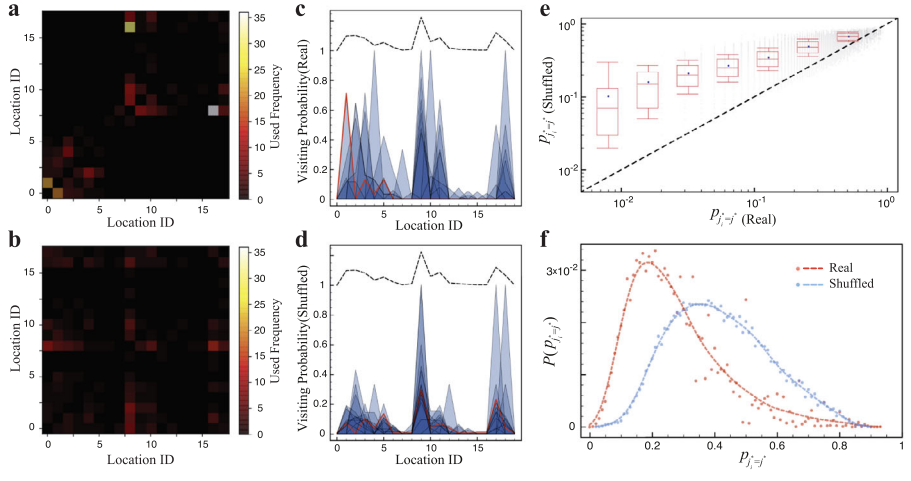

Figure 3 Origin-dependent mobility behavior. Heat maps which show the matrices of the travel frequency of a typical user from one location to another in (**a**) the real data, and (**b**) the shuffled data of the selected typical user. The location visitation probability in (**c**) the real data and (**d**) the shuffled data by the selected typical user originated from specific locations. As an example, the red curves in (**c**) and (**d**) shows the visitation probability distribution of the selected user originated from location 1, while the black curves show the visitation probability distribution aggregated from all starting locations. In (**e**) and (**f**), we show the probability of locations from which the most frequent locations to be next visited is the same as the overall most frequently visited locations (i.e. $p_{j_i^*=j^*}$). The probability $p_{j_i^*=j^*}$ is calculated for each user in both the real data and the shuffled data. (**e**) shows the scatter plot of $p_{j_i^*=j^*}$, indicating that in the real data the most likely locations to be next visited from many locations are different from the overall most frequently locations to be visited. (**f**) shows the distribution of $p_{j_i^*=j^*}$ in real data and the shuffled data

ferent locations, suggesting that the next location that a user visits is not always the most frequently visited ones, but instead strongly depends on his present location. Similarly, we show the visitation probability distribution for each starting location in the shuffled data in Fig. 3d, of which the peaks of the blue curves are consistent with those of the black dashed lines. The comparison between Figs. 3c and 3d shows that in the real data, users' preference on the locations to be visited are dependent on their current location.

A more quantitative analysis can be made by computing the probability that the most frequently visited location j_i^* from location i is consistent with the overall most frequently visited location j^*, i.e. $p_{j_i^*=j^*}$. Figure 3e shows the scatter plot and the bin average of $p_{j_i^*=j^*}$ for each user in the real and the shuffled data. Figure 3f shows the distribution of $p_{j_i^*=j^*}$ for all users in the real and the shuffled data. Both figures show that $p_{j_i^*=j^*}$ is smaller in the real data than that in the shuffled data, again suggesting the origin-dependent preference on the locations to be visited.

Data-integrated Models. With the comprehensive cell-phone position dataset and based on our previous findings, we go on to examine the essential mechanisms underlying human mobility patterns. To achieve the goal, we plug various empirical quantities such as the popularity of locations and the frequency of transition between locations into existing human mobility models, and compare the emergent behavior from the models with empirical results.

We first start with the simplest *preferential return* model of which the probability for an individual to visit a location is proportional to the frequency the location was visited in the past [2]. We can thus write down the transition probability $p_{\alpha:i\to j}(t)$ of an individual α to travel from a location i to a location j at time t to be

$$p^{\mathrm{IPR}}_{\alpha:i\to j}(t) \propto f_{\alpha:j}(t), \tag{1}$$

where $f_{\alpha:j}(t)$ is the empirical frequency that a location j is visited by an individual α before time t. We call the above the *individual preferential return* (IPR) mechanism. A simulated trajectory with (1) to be the transition probability is shown in Fig. 4b, again compared with the real empirical trajectory shown in Fig. 4a. As we can see, many transitions absent in the empirical data are found in the simulated results. Furthermore, we consider a metric $d_\alpha^{\mathrm{total}}$ to examine statistically the validity of this model. We use $d_\alpha^{\mathrm{total}}$ because it is a geographic-aware metric which captures even small inaccurate predictions of paths in the users' travel trajectory. As shown in the scatter plots of $d_\alpha^{\mathrm{total}}$ in Fig. 4g, other than a specific individual, many of the simulated trajectories are longer than their counterparts in the empirical data, which may be a result of the transitions between more distant locations in simulations as in Fig. 4b. These results imply that the IPR mechanism is insufficient to explain human mobility patterns. Since the data of IPR are independent of origin, one may expect that origin-dependent transitions are indeed crucial in explaining mobility patterns.

While the preferential return model is over-simplified in explaining human movement, we then explore the significance of origin-dependent transitions in explaining mobility patterns. Since the individual frequency of transition between two locations is difficult to be modeled, many existing studies only utilize the average transition frequency over the population. Related models for predicting the average transition frequency over the population include the gravity model [40], radiation model [6], population-weighted opportunity model [22] and so on. We call this the *population preferential transition* (PPT)

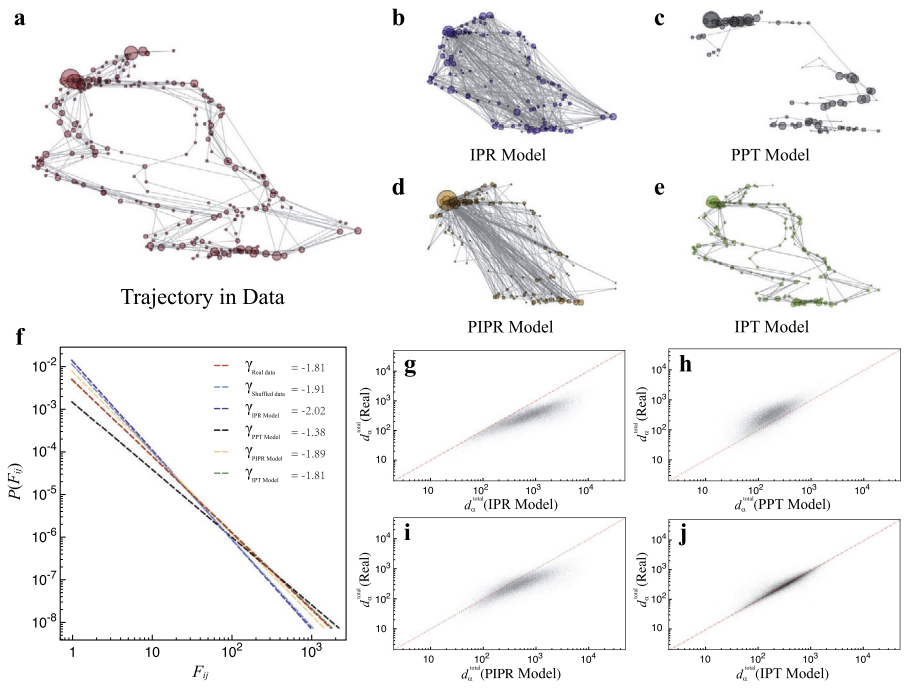

Figure 4 Quantitative comparison of the IPT model with three representative models. (**a**) The travel trajectory of a typical user in (**a**) the real data, and those simulated by (**b**) the IPR model, (**c**) the PPT model, (**d**) the PIPR model, and (**e**) the IPT model. For all the models, the relevant quantities are extracted from the real data of the selected user and used as the initial condition in the simulation. Here each node is a location visited by the user, with node size proportional to its visitation frequency generated by the corresponding model. (**f**) Comparison of the distributions of flux between location pairs, $P(F_{ij})$, in the real and the shuffled data as well as the simulated trajectories from different models. The results of the fitted exponents suggest that the IPT model can best reproduce the flux distribution in real data. The scatter plots of the total traveled distance d_α^{total} of each user α between the real data and the simulated data by (**g**) the IPR model, (**h**) the PPT model, (**i**) the PIPR model, and (**j**) the IPT model

mechanism, of which the transition probability $p_{\alpha:i\to j}(t)$ is given by

$$p^{\text{PPT}}_{\alpha:i\to j}(t) \propto f_{i\to j}(t), \tag{2}$$

where $f_{i\to j}(t)$ is the empirical frequency of which the population travel from location i to j before time t. As shown in Fig. 4c, the trajectory of this specific individual is dominated by paths which connect between near locations, reflecting the average behavior of the population to go to near and attractive locations [6, 22, 40]. This trajectory in Fig. 4c is significantly different from the real trajectory in Fig. 4a. Consistently, we see in Fig. 4h that the simulation underestimates the real total travel distance d_α^{total} for most individuals in the empirical data. These results imply that individuals travel to fulfill specific purposes by which short distance is not the main consideration. Although not surprising, the results suggest that the PPT mechanism is insufficient to explain the individual mobility patterns.

In a recent work [18], a model combining the memory effect and the population-induced competition is proposed to simulate human mobility between locations based only on their population. Basically, individual mobility in this model is driven by both preferential return and collective mobility between locations. In order to test whether this model can generate realistic human mobility at high temporal resolution data, we consider a

population-weighted individual preferential return model (PIPR) combining IPR and PPT, with the transition probability given by

$$p^{\mathrm{PIPR}}_{\alpha:i\to j}(t) \propto f_{\alpha:j} \times f_{i\to j}(t). \tag{3}$$

This model is actually a simplified version of the model proposed in ref. [18], where the collective mobility between locations as predicted by popularity distribution is replaced by the population preferential transition probability. As shown in Figs. 4d and 4i, although the trajectory and the total travel distance are more similar to the empirical data than merely IPR or PPT, they are still different from the real data as it substantially underestimates $d^{\mathrm{total}}_{\alpha}$ in the high temporal resolution human mobility data.

Inspired by the empirical observation in Fig. 3 that people tend to repeatedly transit between a small number of location pairs, we consider here another model based on the first-order Markov process that might explain the driving mechanism in the high temporal resolution human mobility. We call the mechanism the *individual preferential transition* (IPT). In this case, the transition probability $p_{\alpha:i\to j}(t)$ is given by

$$p^{\mathrm{IPT}}_{\alpha:i\to j}(t) \propto f_{\alpha:i\to j}(t), \tag{4}$$

where $f_{\alpha:i\to j}(t)$ is the empirical frequency of which individual α travels from location i to j before time t. As we can see in Fig. 4e, the simulated trajectory resembles the real trajectory shown in Fig. 4a. Other than this specific individual, we see in Fig. 4j that the simulated $d^{\mathrm{total}}_{\alpha}$ of each individual shows a more linear relation with their counterparts in the real data, compared to the above three models (see Figs. 4g, 4h and 4i respectively). These results imply that the IPT mechanism outperforms other factors of preferential return or population competition in capturing human mobility trajectories in high temporal resolution.

When simulating the four models (i.e., IPR, PPT, PIPR, IPT, see Table 1), we draw the initial configurations of these models from the real data. Specifically, $f_{\alpha:j}(t)$ in IPR, $f_{i\to j}(t)$ in PPT, $f_{\alpha:i\to j}(t)$ in IPT are set to be the values extracted from the empirical data. The vectors of $f_{\alpha:j}(t)$ for each user α in IPR and the matrices of $f_{\alpha:i\to j}(t)$ for each user α in IPT are then updated during the simulation. In the IPT model, $f_{\alpha:i\to j}(t)$ increases by 1 if individual α travels from location i to j during the simulation. Similarly, in the IPR model, $f_{\alpha:j}(t)$ increases by 1 if individual α visits location j during the simulation. We stop the simulation for an individual α after he/she finishes the same number of travels as in his/her real data for 14 days.

A remarkable advantage of the state-of-the-art human mobility models is that they can reproduce collective human mobility by aggregating simulated individual mobility trajectories [18]. One important metric that is usually used to examine this feature is the distribution $P(F_{ij})$ of the flux between locations. Figure 4f presents respectively the fitted curves

Table 1 Acronyms for the studied human mobility models

Acronyms	Full name	Mechanism
IPR	Individual Preferential Return	Eq. (1)
PPT	Population Preferential Transition	Eq. (2)
PIPR	Population-weighted Individual Preferential Return	Eq. (3)
IPT	Individual Preferential Transition	Eq. (4)

of the power-law flux distribution generated by IPR, PPT, PIPR and IPT models (See the original distributions in Fig. S8 in SI). We compare these fits with that of the real data (in high resolution, stay duration threshold $T = 3$ mins) and the shuffled data. The exponents with relative errors are: -1.81 ± 0.03 (Real data), -1.91 ± 0.04 (Shuffled data), -2.02 ± 0.04 (IPR), -1.38 ± 0.02 (PPT), -1.89 ± 0.03 (PIPR), -1.81 ± 0.03 (IPT). The relative errors are the difference between the maximal and minimal exponents obtained by varying the fitting curves within the 95% confidence interval (see Fig. S8 in SI for the visualization of the zone of the 95% confidential intervals). As we can see, the exponent generated by the PIPR model is very close to that of the real data. However, the exponent generated by the IPT model is close to that of the real data, suggesting that IPT can best reproduce the real flux distribution. In addition, the difference between these exponents are much larger than the relative errors, supporting that the distribution generated by IPT is closest to the real data.

To understand more comprehensively the difference between the IPT and IPR models, we study several additional metrics, with the results summarized in SI note 5. At individual level, we examine three other metrics including the number n_α^{pair} of transited location pairs, the variance Var_α of the transited pairs' usage frequency, and the distance d_α^{loop} of maximum loop, as presented in Fig. S9. While IPR can reproduce the number of transited location pairs similar to that in the real data, it underestimates Var_α, and overestimates d_α^{loop}. In Fig. S10, we study another metric at the collective level, namely the distribution $F(k_i)$ of the number of different locations that users travel to starting from location i. A longer tail generated by the IPR model indicates that IPR would overestimate the number of different locations that users travel to originated from a specific location. IPT outperforms IPR in reproducing these metrics at both individual and collective levels.

We finally simulate respectively the IPR and the IPT models in a finite space of M locations with no initial memory, in which $N = 6 \times 10^4$ individuals move s steps (with M and s randomly drawn from [2, 350] and [50, 800] respectively). All $f_{\alpha:i\to j}(t)$ in the IPT model and $f_{\alpha:j}(t)$ in the IPR model for individual α are set to be the same small value initially (i.e., $f_{\alpha:i\to j}(t) = 1$ and $f_{\alpha:j}(t) = 1$ for simplicity) and then updated during the process (see details in SI note 5). The results suggest that IPT outperforms IPR in reproducing the observed mobility patterns in the real data, even without the initial memory from the empirical data, see Fig. S11. Specifically, the simulated data from IPT has a smaller number of unique paths and a larger variance of the usage frequency of paths than the corresponding shuffled data, indicating that individuals in IPT tend to use a small number of paths repeatedly. Taken together, the IPT model, integrated with quantities extracted from the comprehensive cell-phone position dataset, can well reproduce human mobility patterns with high temporal distribution that other models fail to capture.

4 Discussion

To summarize, we presented a comprehensive study of human mobility patterns in different temporal scales with a large sample of 4G cell phone data where the positions of users are recorded in each second. We construct mobility networks of mobile phone users, and compare real mobility networks with randomly shuffled networks. We find that the shuffled networks overestimate largely the total number of transited location pairs and the total traveled distance at short time-scale. The collective statistics such as the population flux between locations are also overestimated. This is due to the fact that in the high resolution human mobility data individuals exhibit clear preference on transitions between locations, which is determined by the frequency of the routes that have been used before.

We finally study a simple model based on the first-order Markov process (called individual preferential transition) where the preference of users on paths are accumulated in a matrix and users move according to their preferred paths. The model can quantitatively reproduce the empirical travel patterns at both the individual and population levels up to the high temporal resolution of our empirical data.

Promising future directions include improving the model by introducing the decay of the preference on paths with time, which will result in a more realistic model where the frequently used paths of an individual evolve. In addition, one can empirically study the path preference matrix of individuals, which provides clues to various human mobility behaviors such as explorers and returners observed at the population level in the literatures [16]. Other directions include extending the present work to multiple spatial scales across cities or even countries [6, 18]. The ultimate goal is to obtain a universal model that can be applied to explain the individual and collective human mobility patterns at different spatial and temporal scales. From the perspective of applications, one can study the overlap of users' preference in traveling paths in order to understand and suppress traffic congestion. Answering these questions would not only offer a better understanding of the fundamental mechanisms that underpin individual human mobility, but may also substantially improve our ability to predict and control collective traffic flux [41].

Finally, we remark that our findings can be put in the broader context of complex dynamical systems. The power-law distribution of the flow along edges, and similarly, the stay-time distribution were also studied in the preferential behaviour and scaling in diffusive dynamical systems on networks [42]. The mathematical framework of the discrete-time absorbing Markov chain is also connected to the production optimization in economy [43]. We hope that our findings can inspire new observations and new models in these complex dynamical systems.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-021-00261-2.

Additional file 1. Supplementary information (PDF 2.2 MB)

Acknowledgements
Not applicable.

Funding
This work is supported by the National Natural Science Foundation of China (Nos. 61703136 and 61672206), the Natural Science Foundation of Hebei (No. F2020205012), the Natural Science Foundation of Hebei Education Department (No. QN2017088) and the Youth Top Talent Project of Hebei Education Department (NO. BJ2020035). CHY acknowledges the Research Grants Council of Hong Kong Special Administrative Region, China (Project No. EdUHK 28300215, EdUHK 18304316 and EdUHK 18301217).

Availability of data and materials
The raw data are not publicly available to preserve users' privacy under the Mobile Privacy Policy of China. Derived data supporting the findings of this study are available from the corresponding authors upon request. Due to the data security of participants, 14-days trajectory data cannot be shared freely, but are partly available to researchers who sign a confidentiality agreement and meet the criteria for access to confidential data.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AZ and CZ designed the research, CZ performed the experiments, all authors analyzed the results and wrote the manuscript. All authors read and approved the final manuscript.

Author details
[1]College of Computer and Cyber Security, Hebei Normal University, 050024 Shijiazhuang, P.R. China. [2]Hebei Key Laboratory of Network and Information Security, 050024 Shijiazhuang, P.R. China. [3]Hebei Provincial Engineering Research Center for Supply Chain Big Data Analytics & Data Security, 050024 Shijiazhuang, P.R. China. [4]School of Systems Science, Beijing Normal University, 100875 Beijing, P.R. China. [5]Department of Science and Environmental Studies, The Education University of Hong Kong, Hong Kong, P.R. China.

Publisher's Note

Received: 28 January 2020 Accepted: 11 January 2021 Published online: 19 January 2021

References

1. González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453:779
2. Song C, Koren T, Wang P, Barabási AL (2010) Modelling the scaling properties of human mobility. Nat Phys 6:818
3. Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439:462
4. Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, Blondel VD, Tatem AJ (2014) Dynamic population mapping using mobile phone data. Proc Natl Acad Sci USA 111:15888
5. Hu Y, Zhang J, Huan D, Di Z-R (2011) Toward a general understanding of the scaling laws in human and animal mobility. Europhys Lett 96:38006
6. Simini F, González MC, Maritan A, Barabási AL (2012) A universal model for mobility and migration patterns. Nature 484:96
7. Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: universal patterns in human urban mobility. PLoS ONE 7:e37027
8. Lenormand M, Huet S, Gargiulo F, Deffuant G (2012) A universal model of commuting networks. PLoS ONE 7:e45985
9. Goh S, Lee K, Park JS, Choi MY (2012) Modification of the gravity model and application to the metropolitan Seoul subway system. Phys Rev E 86:026102
10. Simini F, Maritan A, Néda Z (2013) Human mobility in a continuum approach. PLoS ONE 8:e60069
11. Hou L, Pan X, Guo Q, Liu J-G (2014) Memory effect of the online user preference. Sci Rep 4:06560
12. Gallotti R, Bazzani A, Rambaldi S, Barthelemy M (2016) A stochastic model of randomly accelerated walkers for human mobility. Nat Commun 7:12600
13. Szell M, Sinatra R, Petri G, Thurner S, Latora V (2012) Understanding mobility in a social petri dish. Sci Rep 2:457
14. Zhao Z-D, Huang Z-G, Huang L, Liu H, Lai Y-C (2014) Scaling and correlation of human movements in cyber and physical spaces. Phys Rev E 90:050802(R)
15. Zhao YM, Zeng A, Yan XY, Wang WX, Lai YC (2016) Unified underpinning of human mobility in the real world and cyberspace. New J Phys 18:053025
16. Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási AL (2015) Returners and explorers dichotomy in human mobility. Nat Commun 6:8166
17. Alessandretti L, Sapiezynski P, Sekara V, Lehmann S, Baronchelli A (2018) Evidence for a conserved quantity in human mobility. Nat Hum Behav 2:485
18. Yan XY, Wang WX, Gao ZY, Lai YC (2017) Universal model of individual and population mobility on diverse spatial scales. Nat Commun 8:1639
19. Lu X, Bengtsson L, Holme P (2012) Predictability of population displacement after the 2010 Haiti earthquake. Proc Natl Acad Sci USA 109:11576
20. Li X, Xu H, Chen J, Chen Q, Zhang J, Di Z (2016) Characterizing the international migration barriers with a probabilistic multilateral migration model. Sci Rep 6:32522
21. Ren Y, Ercsey-Ravasz M, Wang P, González MC, Toroczkai Z (2014) Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. Nat Commun 5:5347
22. Yan XY, Zhao C, Fan Y, Di Z, Wang WX (2014) Universal predictability of mobility patterns in cities. J R Soc Interface 11:20140834
23. Hasan S, Schneider CM, Ukkusuri SV, González MC (2013) Spatiotemporal patterns of urban human mobility. J Stat Phys 151:304
24. Geng W, Yang G (2017) Partial correlation between spatial and temporal regularities of human mobility. Sci Rep 7:6249
25. Belik V, Geisel T, Brockmann D (2011) Natural human mobility patterns and spatial spread of infectious diseases. Phys Rev X 1:011001
26. Bengtsson L, Lu X, Thorson A, Garfield R, Von Schreeb J (2011) Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. PLoS Med 8:e1001083
27. Vazifeh MM, Santi P, Resta G, Strogatz SH, Ratti C (2018) Addressing the minimum fleet problem in on-demand urban mobility. Nature 557:534
28. Jiang S, Yang Y, Gupta S, Veneziano D, Athavale S, González MC (2016) The TimeGeo modeling framework for urban mobility without travel surveys. Proc Natl Acad Sci USA 113:E5370
29. Lee M, Barbosa H, Youn H, Holme P, Ghoshal G (2017) Morphology of travel routes and the organization of cities. Nat Commun 8:2229
30. Alexander L, Jiang S, Murga M, González MC (2015) Origin-destination trips by purpose and time of day inferred from mobile phone data. Transp Res, Part C, Emerg Technol 58:240–250
31. Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. EPJ Data Sci 4:10
32. Zhao Z et al (2019) The effect of temporal sampling intervals on typical human mobility indicators obtained from mobile phone location data. Int J Geogr Inf Sci 33:1471

33. Toole JL, Colak S, Sturt B, Alexander LP, Evsukoff A, González MC (2015) The path most traveled: travel demand estimation using big data resources. Transp Res, Part C, Emerg Technol 58:162–177
34. Çolak S, Lima A, González MC (2016) Understanding congested travel in urban areas. Nat Commun 7:10793
35. Lima A, Stanojevic R, Papagiannaki D, Rodriguez P, González MC (2016) Understanding individual routing behaviour. J R Soc Interface 13:20160021
36. Yan XY, Han XP, Wang BH, Zhou T (2013) Diversity of individual mobility patterns and emergence of aggregated scaling laws. Sci Rep 3:2678
37. Wang X, Fan T, Li W, Yu R, Bullock D, Wu B, Tremont P (2016) Speed variation during peak and off-peak hours on urban arterials in Shanghai. Transp Res, Part C, Emerg Technol 67:84
38. Jo H-H, Karsai M, Karikoski J, Kaski K (2012) Spatiotemporal correlations of handset-based service usages. EPJ Data Sci 1:10
39. Schneider CM, Belik V, Couronné T, Smoreda Z, González MC (2013) Unravelling daily human mobility motifs. J R Soc Interface 10:20130246
40. Erlander S, Stewart NF (1990) The gravity model in transportation analysis: theory and extensions. VSP
41. Pollock K (2016) Urban physics. Nature 531:S64–S64
42. Kujawski B, Tadic B, Rodgers GJ (2007) Preferential behaviour and scaling in diffusive dynamics on networks. New J Phys 9:154
43. Kostoska O, Stojkoski V, Kocarev L (2020) On the structure of the world economy: an absorbing Markov chain approach. Entropy 22:482