TimeGeo: modeling urban mobility without travel surveys

Shan Jiang[a,1], **Yingxiang Yang**[a,1], **Siddharth Gupta**[a], **Daniele Veneziano**[a], **Shounak Athavale**[b], **and Marta C. González**[a,c,2]

[a]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Ford Motor Company, Dearborn, MI 48126; [c]Center for Advanced Urbanism, Massachusetts Institute of Technology, Cambridge, MA 02139

This manuscript was compiled on April 6, 2016

Well established fine-scale urban mobility models today depend on detailed but cumbersome and expensive travel surveys for their calibration. Not much is known, however, about the set of mechanisms needed to generate complete mobility profiles if only using passive data-sets with mostly sparse traces of individuals. In this study, we present a novel mechanistic modeling framework (TimeGeo) that effectively generates urban mobility patterns with resolution of ten minutes and hundred of meters. It ties together the inference of *home* and *work* activity locations from data, with the modeling of flexible activities (e.g., *other*) in space and time. The temporal choices are captured by only three features: the *weekly home-based tour number*, the *dwell* rate, and the *burst* rate. These combined generate for each individual: (i) stay duration of activities, (ii) number of visited locations per day, and (iii) daily mobility networks. These parameters capture how an individual deviates from the circadian rhythm of the population, and generate the wide spectrum of empirically observed mobility behaviors. The spatial choices of visited locations are modeled by a rank-based exploration and preferential return (r-EPR) mechanism that incorporates space in the EPR model. Finally, we show that a hierarchical multiplicative cascade method can measure the interaction between land use and generation of trips. In this way, urban structure is directly related to the observed distance of travels. This novel framework allows us to fully embrace the massive amount of individual data generated by information and communication technologies (ICTs) worldwide to comprehensively model urban mobility without travel surveys.

human mobility | urban model | mobile phone data

Our ability to correctly model urban daily activities for traffic control, energy consumption and urban planning [1, 2] have critical impacts on people's quality of life and the everyday functioning of our cities. To inform policy making of important projects such as planning a new metro line and managing the traffic demand during big events, or to prepare for emergencies, we need reliable models of urban travel demand. These are models with high resolution that simulate individual mobility for an entire region [3, 4]. Traditionally, inputs for such models are based on census and household travel surveys. These surveys collect information about individuals (socioeconomic, demographic, *etc.*), their household (size, structure, relationships), and their journeys on a given day. Nonetheless, the high costs of gathering the surveys put severe limits on their sample sizes and frequencies. In most cases, they capture only 1% of the urban household population once in a decade with information of only one or few days per individual. The low sampling rate has made it very costly to infer choices of the entire urban population [3, 5–7].

More recent studies try to learn about human behavior in cities by using data collected from location-aware technologies, instead of manual surveys, to infer the preferences in travel decisions that are needed to calibrate existing choice modeling frameworks [8–10]. The problem, however, is that the geotagged data available from communication technologies, in the massive and low cost form, cannot inform us about the detailed activity choices of their users, making most of the data useless for meaningful urban scale mobility models. In order to make the best use of the massive and passive data, a fundamental paradigm shift is needed to model urban mobility and enhance new opportunities emerging through urban computing [11]. This is our goal with TimeGeo, a modeling framework that extracts individual features and key mechanisms needed to effectively generate complete urban mobility profiles from the sparse and incomplete information available in telecommunication activities.

Mobile phones are the prevalent communication tools of the twenty-first century, with the worldwide coverage up to 96% of the population [12]. The call detailed records (CDRs), managed by mobile phone service providers for billing purposes, contain information in the form of geo-located traces of users across the globe. Mobile phone data have been useful so far to improve our knowledge on human mobility at unprecedented scale, informing us about the frequency and the number of visited locations over long term observations [13–18], daily mobility networks of individuals [15, 19], and the distribution of trip distances [13, 15, 17, 20–22]. Due to the sparse nature of mobile phone usage, these data sources have sampling biases and do not provide complete journeys in space and time for each individual [9]. Nonetheless, it has been possible to extract and characterize from phone data where each

Significance Statement

Individual mobility models are important in a wide range of application areas. Current mainstream urban mobility models require socio-demographic information from costly manual surveys, which are in small sample sizes and updated in low frequency. In this study, we propose a novel individual mobility modeling framework, TimeGeo, that extracts required features from ubiquitous, passive, and sparse digital traces in the ICT era. The model is able to generate individual trajectories in high spatial-temporal resolutions, with interpretable mechanisms and parameters capturing heterogeneous individual travel choices. The modeling framework can flexibly adapt to input data with different resolutions, and be further extended for various modeling purposes.

S.J., Y.Y., and M.C.G. designed research; S.J., Y.Y., S.G., D.V., S.A., and M.C.G. performed research; S.J., Y.Y., and S.G. analyzed data; S.J., Y.Y., D.V.,and M.C.G. wrote the paper.

[1]S.J. and Y.Y. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: martag@mit.edu

individual may stay or pass by, and then infer the types of activities that they engage in at various urban locations depending on the time of their visits [23]. By labeling visited location types for individual users as *home*, *work*, or *other*, representative traffic origin-destination (OD) matrices for an average day and by time of day can be generated [24, 25]. They are aggregated estimates of person-trips between pairs of ODs within few hours, and these results have been successfully validated in various cities against existing travel demand models that required expensive surveys for calibration [24, 25].

A fundamental question still remains on how to perform a spatiotemporal mapping of raw mobile phone data to establish models of travel demand with high spatiotemporal resolution, through which individuals' disaggregated daily journeys can be generated. In the current literature that analyzes sparse geotagged data, the daily temporal behavior of human mobility is either not modeled or oversimplified [16, 26]. For example, previous studies on human dynamics do not explicitly model individual temporal choices, but randomly draw parameters such as waiting time or the number of activities in each active period from aggregated distributions measured from data [14, 15, 27]. The model in [19] introduces time dependency in travel and tendency to arrange short out-of-home activities in consecutive sequences (i.e., bursts of activities) [27–31], but the stay duration at flexible (*other*) locations is fixed. Furthermore, it does not incorporate spatial choices or the heterogeneity of individual behavior.

To realistically model individual mobility in cities at both micro- and macro-level, it is necessary to understand the essential features of a population distribute in space at different times. Here we show that these features can be extracted from big data sources. We present spatiotemporal patterns of individual daily mobility that can be generated by a coherent framework of mechanisms. This paper discusses the first comprehensive process of converting sparse mobility traces into daily trajectories in temporal resolution of ten minutes and spatial accuracy of a few hundred meter radius, with interpretable probabilistic mechanisms. Instead of using social-demographic information to calibrate the set of detailed decisions involved in activity choices—as required by mainstream transportation modeling approaches, the framework consists of directly measurable parameters discovered from passive data. It represents a needed paradigm shift to model individual daily trajectories in cities, adapted to ubiquitously available sparse digital traces of individuals. The results are high resolution travel diaries for a large sample of users based on their ICT data in the urban context. The presented set of parameters can be further refined as more information becomes available at the individual level.

Activity extraction from mobile phone data

To demonstrate the mechanistic modeling framework, we analyze a CDR data set of 1.92 million anonymous mobile phone users for a period of 6 weeks in the Greater Boston area. To have a control experiment, we also examine a donated set of self-collected mobile phone traces of a graduate student in the same region over a course of 14 months, recorded by a smartphone application. When an individual anchors at a location to conduct an activity, it is defined as a *stay*. We apply the stay extraction method discussed in the literature [23] to both data sets. We filter out signal jumps as well as pass-by records when mobile phone users were traveling. For each user, based on the start time and frequency of visits to each stay location, we infer the stay location type as *home* (H), *work* (W), or *other* (O).

We are able to identify *home* locations for 1.44 million users which is 75% of our initial user base. Next, we filter users who have more than 50 total stays and at least 10 home stays in the observation period. These are identified as *active* users and are used to extract the various parameters of TimeGeo (as explained in detail in the next sections). These active users can be labeled as commuters (133,448 individuals) who have journey-to-work trips, and non-commuters (43,606 individuals) who have no journey-to-work trips.

Fig. 1 illustrates the pipeline of extracting stays, labeling activity types, and deriving individual mobility features from raw mobile phone data for each of three demonstrated days. Fig. 1 (a-c) show the raw cell phone records (in blue for 14 months, and in purple for each day), and the extracted stay locations of the individual (in red). Fig. 1 (d-f) show that for active users the extracted stays in each day define a daily journey (usually starting and ending at home). A trip is made when a user changes stay locations. The time-bar shows the start time and duration for each stay, and activity types are color-coded.

Generating mechanisms of individual mobility

The modeling framework of TimeGeo is presented in Fig. 2 (a). It integrates the temporal and spatial choice mechanisms of human mobility. We assume that for an individual agent, her *work* activity has a fixed location, start time, and duration; her *home* activity is fixed in terms of location but flexible with start time and duration; her *other* activity is flexible with regard to location, start time, and duration. The presented framework aims to model the flexible spatial and temporal mobility choices, whereas the schedule of the fixed activity (i.e., work) is assumed as predetermined (see SI Appendix section 2 for details). We divide each day of a week into 144 discrete intervals of 10 minutes (i.e., 1008 time-intervals in a week). For each time interval t within a week, an individual first decides to *stay* or *move*. If she chooses to move, she then decides where to go. We improve from previous human mobility models [14] by generating spatio-temporal patterns while introducing individual-specific mobility parameters, namely: a weekly home-based tour number, a dwell rate, and a burst rate (explicitly defined later). These parameters capture the heterogeneity of individual daily mobility observed in the passive digital traces. Nevertheless, due to the limited observation period of the CDR data used in this study, some parameters cannot be extracted at the individual level. These global parameters measure the preferential return and exploration rates, and the rank selection probability. As large scale data with higher frequency (e.g., GPS traces) and longer observation periods (e.g., many months) become available, these global parameters could be measured at the individual level as well.

Temporal choices. To uncover the key generating mechanisms needed to reproduce individual daily trajectories, we propose a time-inhomogeneous Markov model with three individual-specific parameters—*weekly home-based tour number* (n_w), *dwell rate* (β_1), and *burst rate* (β_2)—to capture individual

circadian propensity to travel [16, 19, 26] and likelihood of arranging short activities in consecutive sequences [27–31]. As *work* activity is assumed to have fixed start time and duration, we consider two Markov states: *home* and *other*. *Home* is considered as a *less-active* state, since the average stay duration at *home* is significantly longer than that at *other* states where people are more *active* (*i.e.,* likely to travel).

When an individual l is at *home*, her individual travel circadian rhythm is defined as $n_w P(t)$, representing her likelihood of making a trip originated from *home* in a time-interval t of a week. The *weekly home-based tour number* n_w counts the total number of trips that an individual l initiated from *home* to *other* places. $P(t)$ is the global travel circadian rhythm of the population in an average week. We differentiate $P(t)$ for commuters and non-commuters (see SI Appendix section 3.1). For non-commuters, $P(t)$ is measured as the fraction of all user-trips in the time interval t of the week for the population (*i.e.* $\sum_{t=1}^{1008} P(t) = 1,\ t = 1, 2, ..., 1008$), capturing the expected variation of travel in different time of the week (shown in Fig. 2 (b)). For commuters, since *work* is modeled as a fixed activity, $P(t)$ does not include trips to or from *work*. The product of the two, $n_w P(t)$, less than 1, defines the individual travel probability at a specific time interval (t) while she is at home.

To model an individual's propensity to travel from an *other* (out-of-home) state, we introduce a *dwell rate* β_1 which measures how much more *active* (or likely to travel) the person is at an *other* state compared to *home*. The probability of traveling when an individual is at an *other* state is defined as $\beta_1 n_w P(t)$. By capturing individual propensity to move from an *other* state, $\beta_1 n_w$ controls the stay duration Δt for flexible activities. The higher the product $\beta_1 n_w$, the more likely the person will choose to move and thus the shorter duration Δt she will stay at *other* locations.

Next, if an individual is already out of home and chooses to move at time t, we then model her decision to either go *home* or go to an additional *other* location by introducing a *burst rate* β_2. We measure the probability that the individual travels from an *other* location O_1 to an additional *other* location O_2 as $P(O_1 \to O_2) = \beta_2 n_w P(t)$. It is assumed that for an individual who has decided to move, the probability of visiting an additional *other* location is proportional to $\beta_2 n_w$. The ratio between the two choices of going to an additional *other* location or going *home* can be presented as follows:

$$\frac{P(O_1 \to O_2)}{P(O_1 \to H)} = \frac{\beta_2 n_w P(t)}{1 - \beta_2 n_w P(t)}, \qquad [1]$$

For a given value of $\beta_2 n_w$, when $P(t)$ is high (*e.g.,* in the afternoon), people are more likely to visit additional other locations; when $P(t)$ is low, people are more likely to return home. For a given $P(t)$, the higher the value of $\beta_2 n_w$, the higher probability the individual will keep visiting flexible (*other*) locations, and thus the greater number of daily locations N she will visit.

Compared to previous models that randomly draw the stay duration (or waiting time Δt) or the number of visited locations (N) from aggregated empirical distributions [14, 27], by introducing three individual-specific parameters including *weekly home-based tour number* n_w, *dwell rate* β_1, and *burst rate* β_2, we explicitly model the temporal dynamics of individual mobility. The Markov model framework allows it to be analytically tractable and to derive explicit effects in the resulting stay-duration and daily-location distributions $P(\Delta t)$ and $P(N)$ (see SI Appendix section 6).

Spatial choices. To model the spatial choices of individual mobility, we propose a rank-based exploration and preferential return (r-EPR) model by incorporating a rank-based selection of new locations to the original EPR model [14]. The EPR model explains well the differences in the frequency of visits of each location [13–18, 32]. For each movement, an individual decides either to explore a new location with probability P_{new}, or return to a previously visited location with probability $1 - P_{new}$. The exploration probability $P_{new} = \rho S^{-\gamma}$ captures a decreasing propensity to visit new locations as the number of previously visited locations (S) increases with time, and effectively captures individual mobility choices between explorations and returns. If the individual decides to return to previously visited locations, she chooses a specific location i with probability P_i defined as the visitation frequency of location i [14]. Fig. 1 (g-i) illustrate P_i with different circle sizes, using the volunteered student's location records as an example. In each sub-figure, we label the visitation frequency of each location up to the current day. We highlight locations visited in the current day in the foreground and show the previously visited ones in the background.

If the individual decides to explore a new location, she needs to choose a destination from a large number of possible alternatives. One limitation of the original EPR model proposed in [14] is its lack of a mechanism for the new-location selection. To select a new location, the original EPR model randomly draws the exploration jump-size (Δr) from a global empirical distribution. To model the exploration mechanism more sensible to the urban structure, in this study, we incorporate a rank-based selection mechanism for newly explored locations (*i.e., r-EPR* model).

Our selection mechanism gives a rank k to each alternative destination based on their distances to the trip origin [33–36]. Among all potential new destinations, the one closest to the current location is of *k=1*, the second closest *k=2*, *etc.* The empirical probability of selecting the k-th location as a destination is quantified as $P(k) \sim k^{-\alpha}$, the same form has been measured in various studies that analyze aggregated trips between locations for both commuting and non-commuting trips [33–36]. For an individual to select an exploration-destination, we measure $P(k)$ aggregating all users' destinations. Figs. 1 (j-l) illustrate probabilities of selecting different destinations (with higher ranks in red and lower ranks in blue). Each dot represents a location for an *other* activity extracted from the CDR data. The height of the dot on the z axis represents the dot density at the location.

Because the observation period of the empirical data in this study is six weeks, most users have a limited number of exploration-trips, making it difficult to estimate the spatial parameters of $P(k)$ at the individual level. Given more abundant data, this distribution could be estimated at the individual level as well.

The role of land-use on travel distance

Different spatial patterns of cities imply different geographical advantages to urban functioning [37]. TimeGeo takes the spatial distribution of locations (*e.g.*, observed from the CDR

data) as an input. To explain and quantify the influence of land-use on travel, we propose a hierarchical multiplicative cascade framework of analysis. It allows scenario tests on how changes in land-use patterns will affect individual travel. It can generate different scenarios of urban structure (*i.e.,* spatial distribution of *home* and *other* activities).

Fig. 3 (a-d) show the distribution of different types of locations (*home* and *other*) extracted from the mobile phone data set at two scales: At a scale with larger grids, *home* and *other* locations are mixed spatially, showing high spatial correlations. At a scale with smaller grids, the separation between *home* and *other* types of land-use becomes clear [35]. The intuition behind this phenomenon is that at a scale with smaller grids (*e.g.*, similar to the census block level), land use is often separated— meaning that residential land use is separated from non-residential one; while at a scale with larger grids (*e.g.*, at the district, town, or regional level), residential and non-residential land use mix together. A hierarchical multiplicative cascade divides an area of interest into grids with different granularity and quantifies the spatial correlation of each type of land uses at different scales.

The current framework integrates the two features that influence the spatial choices of exploration to *other* locations. These are (i) the spatial distribution of activity locations, and (ii) the rank-based location-selection mechanism (illustrated in Fig. 1 (j-l)). By characterizing the spatial distributions of population and facilities at various scales, here we formalize how these two features influence the observed trip-distance distribution.

To quantitatively represent *home* to *other* ($H-O$) trip distance, we denote *home* locations as the demand side D, and *other* locations as the supply side S. The entire region of interest is Ω_0 (taken as a unit square, shown in Fig. 3(e)). We progressively partition Ω_0 into 4^1, 4^2, ..., 4^n square-tiles with side length $2^{-1}, 2^{-2}, ..., 2^{-n}$. Each time a mother-tile Ω_{i-1} (at resolution level $i-1$) is partitioned into 4 daughter-tiles Ω_i (at resolution level i). Then the probability that a trip goes outside its origin-tile at resolution level i, $P_>(i)$, can be expressed as

$$P_>(i) = \int_1^M P_>(k) f_{S_{i,trip}}(k) dk \qquad [2]$$

where M is the total number of supplies in the entire region Ω_0; $P_>(k)$ is the probability that the k supplies in the origin-tile are not chosen; $f_{S_{i,trip}}(k)$ is the probability of finding k supplies within the origin tile. The tile exceeding probability $P_>(i)$ at different tile resolutions generates the resulting distribution of trip distances. Equation 2 ties together the rank-based selection mechanism $P_>(k)$ and the geographic distribution of locations $f_{S_{i,trip}}(k)$, which can be calculated as

$$f_{S_{i,trip}}(k) = \int_0^Q f_{D_{i,trip}}(D) f_{S_i|D_i=D}(k) dD \qquad [3]$$

where $f_{D_{i,trip}}(D)$ is the conditional probability that a trip originates in a tile at level i given D demands are in that tile. $f_{S_i|D_i}$ is the conditional probability of supply given demand. Q is the number of demand in the entire study area. In summary, to quantify trip distance through $P_>(i)$, we not only need the distribution of each type (*home* and *other*) of location, but also the correlation between them at different scales. The detailed introduction to the cascade method of analysis can be found in Ref. [38] and in the *Methods* section, the derivation of the resulting trip distance distribution is presented in SI Appendix section 5.

Results

Extracted mobility features from mobile phone data. In this section we show the results for non-commuters. For each individual, the *weekly home-based tour number* n_w is directly extracted from the data. While the β_1 and β_2 parameters are calibrated using the temporal Markov model. The rest of the parameters needed are: $\alpha = 0.86$ for the rank selection probability $P(k) \sim k^{-\alpha}$, and $\rho = 0.6$ and $\gamma = 0.21$ for the preferential return mechanism $P_{new} = \rho S^{-\gamma}$. These three parameters are extracted from the aggregated data of the entire population (Fig. 2(d, e)).

The individual values of β_1 and β_2 values, are obtained by calibrating the Markov model to minimize the following statistic:

$$A(\beta_1, \beta_2) = \int |P_D(\Delta t) - P_M(\Delta t|\beta_1, \beta_2)| d\Delta t + \eta |\bar{N}_D - \bar{N}_M(\beta_1, \beta_2)| \qquad [4]$$

where $P_D(\Delta t)$ and $P_M(\Delta t|\beta_1, \beta_2)$ are the distributions of the individual empirical and modeled stay-duration, respectively. Scalar values $\bar{N}_D$ and $\bar{N}_M(\beta_1, \beta_2)$ are the average daily number of visited locations measured from the individual's empirical data and from the model-simulation, respectively. The difference between $\bar{N}_D$ and n_w is that $\bar{N}_D$ counts all trips while n_w only counts trips starting at home. Meta-parameter $\eta = 0.035$ controls the weight between the two components. Since $A(\beta_1, \beta_2)$ is a non-convex function, discrete β_1 and β_2 values are used ($\beta_1 = 1, 2, 3, ..., 20, \beta_2 = 1, 6, 11, ..., 101$) to estimate the (β_1, β_2) pair that minimizes $A(\beta_1, \beta_2)$ for each person. The empirical results of $n_w\beta_1$, $n_w\beta_2$ and n_w for all the individuals is presented in Fig. 2(c). The median value of n_w, $n_w\beta_1$ and $n_w\beta_2$ for non-commuters are 7.4, 34.2, and 355.6 respectively. Median *dwell* rate $\beta_1 = 4.6$, suggesting that when people are not at home, they are on average 4.6 times more likely to travel.

Simulated mobility features. Taking the featured parameters measured directly from active users of the mobile phone data set, TimeGeo can generate realistic individual daily trajectories over a long time period at the urban scale.

We first use the student volunteer's 14-month mobile phone records as an example to explain the simulation and interpret the results of TimeGeo. We fix the locations of *home* and *work* (in this case school is identified as *work*) and apply the proposed modeling framework to simulate the spatiotemporal choices of flexible *other* activities and temporal choices of *home* activities. For the student, we computed that his *dwell rate* $\beta_1 = 4$, *burst rate* $\beta_2 = 36$, and *weekly home-based tour number* $n_w = 7$. His burst rate is lower than the population average, reflecting smaller likelihood to conduct consecutive short activities. Fig. 4 (a-c) show three simulated days for the student. The days are predominated by *home-work* trips, with a few trips to other locations. The model is able to capture not only the number of locations visited each day, but also more detailed configuration of daily trip chains. Fig. 4 (d) shows the distribution of the most frequent daily mobility networks, *i.e.*, daily motifs, of the student. We represent unique locations as nodes and trips between locations as edges and

count the motif distribution for days start and end at home. The dominating motif is traveling just between two locations in a day. To show the infrequent motifs clearer, we present the percentage in log scale.

A key value of TimeGeo is to use ICT records to generate individual trajectories from discovered mobility features at the urban scale. In Fig. 4 (d-f), we illustrate a user with very sparse data. She only had 4 distinct locations in 30 days and we simulate her complete daily trajectories in space and time. We select two different sets of β_1, β_2, and n_w from the joint distribution shown in Fig. 2 (c) to generate two synthetic realizations of the user. Fig. 4 (e-f) show the two resulting profiles of simulated journeys of the same sparse user and Fig. 4 (h) shows the distinct motif distributions.

The importance of the individual features extracted from data (Fig. 2 (c)) lies in its ability to capture diverse travel behaviors observed in the population. Fig. 5 (a-d) compare mobility patterns for different individual profiles. The individual 1 and 2 represent two extreme cases: one travels more frequently (shown in squares, $n_w = 10.86$, $\beta_1 = 6$, $\beta_2 = 41$) and the other travels less frequently (shown in circles, $n_w = 5.51$, $\beta_1 = 1$, $\beta_2 = 36$). As a comparison we also present the average case —a simulation using median values of the parameters n_w, β_1, and β_2. Fig. 5 (a, b) show that these three individuals have distinct $P(\Delta t)$ and $P(N)$ distributions. The less frequent traveler has significantly longer stay duration and visits fewer locations per day. To quantify the differences between empirical distributions of data and the model simulation, we employ the Kolmogorov–Smirnov (KS) test. The KS statistic between empirical and simulated $P(\Delta t)$ for the two extreme individuals are 0.12 and 0.11 respectively. If we compare their empirical data with the *average* case, the KS statistic increases to 0.25 and 0.20 respectively. Similarly, for these two individuals, the KS statistic for $P(N)$ are 0.05 and 0.12. When comparing with the *average* case, the KS statistic increases to 0.40 and 0.50 respectively. It confirms the importance of including individual-specific parameters to model temporal choices. Interestingly, although these three cases share similar location visitation frequency $f(L)$, the simulation result fits less well in trip distance distribution $P(\Delta r)$ (as shown in Fig. 5 (c, d)). The KS statistic for $P(\Delta r)$ are 0.64 and 0.49. One reason is that the model uses global parameter values for spatial choices. On the other hand, as is shown by the aggregated trip distribution in Fig. 5 (h), the proposed model overestimates long-distance trips because although trip distance (rank of each location) is an influential factor when selecting new visitation locations, for return-trips, distance is not a decision factor. Therefore, compared to empirical data, the proposed model has a higher probability of visiting far but frequently visited locations. With data of high frequency and longer observation period available in future studies, machine learning methods can be applied to better learn from choices at individual level when choosing return trips for improvement of our proposed modeling framework.

Fig. 5 (e-h) compare aggregated mobility features extracted from data and simulation for all the active non-commuters. These results show that to reproduce individual mobility patterns realistically, it is critical to incorporate each of the mechanisms proposed in the current modeling framework. Namely, the weekly home-based tour number, dwell rate, burst rate, the rank-based EPR, over the land use profile of the city under consideration. The results on the aggregated daily mobility motif distribution is presented in SI Appendix section 4.2.

For the *dwell rate* (β_1), if $\beta_1 = 1$, *i.e.*, the model does not differentiate the mobility circadian rhythms of *home* or *other* activities. The resulting $P(\Delta t)$ distribution will underestimate trips with short duration, and the KS statistic increases from 0.04 to 0.27. For the $P(N)$ distribution, the KS statistic for the model with and without *the burst rate* β_2 are 0.03 and 0.22 respectively. The bursts of flexible activities, captured by the *dwell and burst rates* β_1 and β_2, ensure realistic distributions of the stay duration $P(\Delta t)$ and the number of daily visited locations $P(N)$. The improved rank-based EPR mechanisms, model the selection of locations. It improves the KS statistic of the trip distance distribution from 0.52 to 0.39. The visitation frequency to the L^{th} most visited location follows $f(L) \sim L^{-1.2\pm0.1}$. In Fig. 3 (f), $P_>(i)$ measures the probability that a generic exploration trip goes outside its origin tile at resolution level i. At the largest 4 tile sizes ($24km$, $12km$, $6km$, $3km$), the cascade is a pure log-normal cascade and $P_>(i)$ can be analytically calculated and the result compares very well with the data. The empirical data, simulation, and analytical calculation all show that 20% of the trips cross the tile with a size of $24km$, and over 60% cross the tile with a size of $3km$.

In SI Appendix section 4.3 we show the results of simulated daily mobility patterns for the population (aged 16 and over) in Metro Boston (3.54 million individuals). By carefully expanding active mobile phone users to the population, we generated 1 weekday mobility trajectories using TimeGeo. Our simulation results show good agreement with the latest travel survey [39] and with two state-of-the-art travel demand models of the Boston Region Metropolitan Planning Organization (MPO) which needed expensive surveys for calibration [40].

Conclusion

We present a mechanistic modeling framework to generate individual daily mobility with fine-resolution at urban scale. Temporally, we introduce the *weekly home-based tour number*, *dwell rate*, and *burst rate* to model the bursts of short flexible activities in activity-chains. This mechanism can reproduce individual distributions of stay duration, number of daily visited locations, and daily mobility motif distribution. Spatially, an improved rank-based EPR model is introduced to explain individual activity location selection choices. Compared to the original EPR model, the ranking mechanism quantifies the likelihood of selecting new destinations in space based on the distribution of facilities around trip origins. Moreover, the covariance of the distributions of population and facilities in a given region are characterized using a hierarchical multiplicative cascade framework of analysis. In this way, we take account of the influence of region-specific spatial structure on individual travel distances. This enables us to perform scenario tests on how changing land use in the city would affect micro-level individual travel behavior and macro-level OD flows.

TimeGeo serves as a general modeling framework of urban trajectories that can be flexibly adapted to different application scenarios using population density and the distributions of facilities in any city. It can be coupled with sparse location data from ICTs that sample the visitation preferences

of actual individuals and can complement or, for some applications, substitute the need for expensive travel surveys for modeling urban travel. The framework is flexible to generate trajectories with various data conditions. The minimum requirement is to have population and facilities distributions. In the current results, the parameters to model exploration and returns (α, ρ and γ) are assumed to be the same across the population, while the temporal mobility rates of an individual are assumed to be independent of the actual location. In future studies, as more data of higher frequency and over longer periods become available, it is possible to further learn from the individual variations of the proposed parameters. It is also interesting to explore the variations of the model parameters across urban areas, and across population groups with different demographics and lifestyles.

Materials and Methods

All study procedures were carried out with Institutional Review Board approval from MIT COUHES (protocol # 1405006399) approved on June 10, 2014. Call Detailed Record (CDR) data was collected by AirSage for billing and operational purposes. The student, who donated his 14-month self-collected mobile phone traces through a smartphone application (OpenPaths), provided informed consent for the research.

Mobile phone data. We extracted activity stay locations of 1.92 million cell phone users from their Call Detailed Records (CDR) in the Greater Boston area during an observation period of 6 weeks. A stay means performing an activity at a location. A stay sequence, or an activity sequence, represents consecutive stays a person made in a period of time (usually a day). A trip is made between consecutive stay locations. These stay locations are also called trip origins and destinations. In the CDR data, a record is made when a user calls, sends text messages, or uses data through the cellular networks. Each record is in the following format: (UserID, longitude, latitude, time). The precision of the location is about 200m to 300m in urban areas. For the voluntarily self-collected mobile phone user example, a record is made every time the smartphone application detects a significant spatial movement. The data is in the same format and similar spatial resolution as the CDR data. The detailed methods to extract stay locations and to label location types (as *home*, *work*, and *other*) are presented in SI Appendix section 1. For the CDR data, the records do not directly correspond to a user's stays—A stay could not be detected if a user does not use his or her cell phone more than once during a stay. Even for cases when more than one cell phone usages were recorded, the stay duration can only be approximated for active phone users. Therefore not all cell phone users have enough records to be measured for basic mobility patterns presented in this study. Meanwhile, we cannot determine if long stays at one location (for over 2 days) are caused by no cell phone usage or actual stay at one location for over 2 days, therefore these stays were removed from the analysis and not captured by the model.

The hierarchical multiplicative cascade model. For any given subregion $\omega \subset \Omega_0$, $D(\omega)$ is the number of trip origins in ω and $S(\omega)$ is the number of trip destinations in ω. We use bivariate random measures $X(\omega) = [D(\omega), S(\omega)]$ to represent the number of demand and supply locations in ω, where X results from a cascade process in which the fluctuations at different spatial scales combine in a multiplicative way. The generation of bivariate $[D, S]$ cascades is illustrated in Fig. 3(c). The demand and supply in a generic i-tile Ω_i are D_i and S_i and the associated measure densities are $D'_i = D_i/|\Omega_i|$ and $S'_i = S_i/|\Omega_i|$. One starts with uniform measure densities D'_0 and S'_0 in Ω_0, then progressively partitions Ω_0 into 4^1, 4^2, ..., 4^n square tiles of side length $2^{-1}, 2^{-2}, ..., 2^{-n}$. The demand and supply densities in the daughter tiles are multiplied by independent realizations of non-negative random factors W_{D_i} and W_{S_i}, with mean value 1. The random vectors $W_i = [W_{D_i}, W_{S_i}], i = 1, 2, ...n$ are the generators of the cascade. While the generators W_i have independent values in different i-tiles, their components W_{D_i} and W_{S_i} in a given i-tile may be dependent. Moreover, the distribution of W_i may vary with the resolution level i. These features provide important modeling flexibility. The measured densities at resolution level $i-1$ and i are related as

$$\begin{bmatrix} D'_i \\ S'_i \end{bmatrix} \stackrel{d}{=} \begin{bmatrix} W_{D_i} & 0 \\ 0 & W_{S_i} \end{bmatrix} \begin{bmatrix} D'_{i-1} \\ S'_{i-1} \end{bmatrix} \quad [5]$$

According to Fig. 3 (a-d), at larger tile sizes almost all tiles are non-empty and the supply and demand have positive correlation. Consequently for small i values (large tile sizes) the generator can be described as joint log-normal variables [38]. If the log generators $\ln(W_{D_i})$ and $\ln(W_{S_i})$ have joint normal distribution with variances $\sigma^2_{W_{D_i}}$ and $\sigma^2_{W_{S_i}}$, mean values $-1/2\sigma^2_{W_{D_i}}$ and $-1/2\sigma^2_{W_{S_i}}$ and correlation coefficient ρ_{LN_i}, then $\ln(D_i)$ and $\ln(S_i)$ have joint normal distribution with mean values m_{D_i} and m_{S_i}, variances $\sigma^2_{D_i}$ and $\sigma^2_{S_i}$ and correlation coefficient ρ_i given by

$$\sigma^2_{D_i} = \sum_{j=1}^{i} \sigma^2_{W_{D_j}}, m_{D_i} = \ln(D_0 4^{-i}) - 1/2\sigma^2_{D_i} \quad [6]$$

$$\sigma^2_{S_i} = \sum_{j=1}^{i} \sigma^2_{W_{S_j}}, m_{S_i} = \ln(S_0 4^{-i}) - 1/2\sigma^2_{S_i} \quad [7]$$

$$\rho_i = \sum_{j=1}^{i} \frac{\rho_{LN_j} \sigma_{W_{D_j}} \sigma_{W_{S_j}}}{\sigma_{D_i} \sigma_{S_i}} \quad [8]$$

Therefore, once we can estimate $\sigma_{W_{D_i}}$, $\sigma_{W_{S_i}}$ and ρ_{LN_j}, the rest of the variables can be calculated.

At smaller tile sizes, empty tiles cannot be ignored and extreme forms of dependence like mutual exclusion may occur. In this case the generator is better modeled as a β cascade, in which a tile is either filled or empty. The generators $W(i) = [W_{D(i)}, W_{S(i)}]$ of a bivariate β cascade have a discrete distribution with probability masses concentrated at four (w_D, w_S) points: mass P_{00} at $(0, 0)$, mass P_{D0} at $(1/P_D, 0)$, mass P_{0S} at $(0, 1/P_S)$, and mass P_{DS} at $(1/P_D, 1/P_S)$. $P_D = P_{D0} + P_{DS}$, $P_S = P_{0S} + P_{DS}$, and $P_{D0} + P_{DS} + P_{0S} + P_{00} = 1$. Thus a tile is either filled or empty. The correlation between the supply and demand is ρ_{β_i}.

ACKNOWLEDGMENTS. We thank Chaoming Song for enlightening discussions during the design of this work. The research reported herein was funded in part by the MIT-Ford alliance, MIT-Philips alliance, the MIT-Brazil program, the MIT-Portugal program, DOT via the program New England UTC 25 and the Center for Complex Engineering Systems (CCES) at KACST.

1. Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4):211–221.
2. Batty M (2013) *The New Science of Cities.* (MIT Press).
3. Nagel K, Beckman RJ, Barrett CL (1999) Transims for urban planning in *6th International Conference on Computers in Urban Planning and Urban Management, Venice, Italy*.
4. Ben-Akiva M, Bierlaire M (1999) Discrete choice methods and their applications to short term travel decisions in *Handbook of transportation science.* (Springer), pp. 5–33.
5. Balmer M et al. (2008) *Agent-based simulation of travel demand: Structure and computational performance of MATSim-T.* (ETH, Eidgenössische Technische Hochschule Zürich, IVT Institut für Verkehrsplanung und Transportsysteme).
6. Arentze T, Timmermans H (2000) *Albatross: a learning based transportation oriented simulation system.* (Eirass Eindhoven).
7. Bowman JL, Ben-Akiva ME (2001) Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice* 35(1):1–28.
8. Danalet A, Tinguely L, Cochon de Lapparent MM, Bierlaire M (2015) Location choice with longitudinal WiFi data, (Lausanne, Switzerland), Technical report.
9. Zilske M, Nagel K (2014) Studying the accuracy of demand generation from mobile phone trajectories with synthetic data. *Procedia Computer Science* 32:802–807.
10. Zilske M, Nagel K (2015) A simulation-based approach for constructing all-day travel chains from mobile phone data. *Procedia Computer Science* 52:468 – 475. The 6th International Conference on Ambient Systems, Networks and Technologies (ANT-2015), the 5th International Conference on Sustainable Energy Information Technology (SEIT-2015).
11. Zheng Y, Capra L, Wolfson O, Yang H (2014) Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5(3):38.

12. Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. *arXiv preprint arXiv:1502.03406*.
13. Gonzalez MC, Hidalgo CA, Barabasi AL (2008) Understanding individual human mobility patterns. *Nature* 453(7196):779–782.
14. Song C, Koren T, Wang P, Barabási AL (2010) Modelling the scaling properties of human mobility. *Nature Physics* 6(10):818–823.
15. Perkins TA et al. (2014) Theory and data for simulating fine-scale human movement in an urban environment. *Journal of The Royal Society Interface* 11(99):20140642.
16. Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. *Science* 327(5968):1018–1021.
17. Hasan S, Schneider CM, Ukkusuri SV, González MC (2013) Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics* 151(1-2):304–318.
18. Toole JL, Herrera-Yaqüe C, Schneider CM, González MC (2015) Coupling human mobility and social ties. *Journal of The Royal Society Interface* 12(105):20141128.
19. Schneider CM, Belik V, Couronné T, Smoreda Z, González MC (2013) Unravelling daily human mobility motifs. *Journal of The Royal Society Interface* 10(84):20130246.
20. Kölbl R, Helbing D (2003) Energy laws in human travel behaviour. *New Journal of Physics* 5(1):48.
21. Balcan D et al. (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences* 106(51):21484–21489.
22. Viswanathan G et al. (1996) Lévy flight search patterns of wandering albatrosses. *Nature* 381(6581):413–415.
23. Jiang S et al. (2013) A review of urban computing for mobile phone traces: Current methods, challenges and opportunities in *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, UrbComp '13. (ACM, New York, NY, USA), pp. 2:1–2:9.
24. Toole JL et al. (2015) The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*.
25. Alexander L, Jiang S, Murga M, González MC (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies* 58:240–250.
26. Jo HH, Karsai M, Kertész J, Kaski K (2012) Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics* 14(1):013055.
27. Malmgren RD, Stouffer DB, Motter AE, Amaral LA (2008) A poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences* 105(47):18153–18158.
28. Vázquez A et al. (2006) Modeling bursts and heavy tails in human dynamics. *Physical Review E* 73(3):036127.
29. Barabasi AL (2005) The origin of bursts and heavy tails in human dynamics. *Nature* 435(7039):207–211.
30. Hidalgo R, César A (2006) Conditions for the emergence of scaling in the inter-event time of uncorrelated and seasonal systems. *Physica A: Statistical Mechanics and its Applications* 369(2):877–883.
31. Karsai M, Kaski K, Barabási AL, Kertész J (2012) Universal features of correlated bursty behaviour. *Scientific reports* 2.
32. Pappalardo L et al. (2015) Returners and explorers dichotomy in human mobility. *Nature communications* 6.
33. Simini F, González MC, Maritan A, Barabási AL (2012) A universal model for mobility and migration patterns. *Nature* 484(7392):96–100.
34. 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(5):e37027.
35. Yang Y, Herrera C, Eagle N, González M (2014) Limits of predictability in commuting flows in the absence of data for calibration. *Scientific reports* 4.
36. Noulas A, Shaw B, Lambiotte R, Mascolo C (2015) Topological properties and temporal dynamics of place networks in urban environments in *Proceedings of the 24th International Conference on World Wide Web Companion*. (International World Wide Web Conferences Steering Committee), pp. 431–441.
37. Batty M (2008) The size, scale, and shape of cities. *Science* 319(5864):769–771.
38. Veneziano D, Gonzalez MC (2010) Trip length distribution under multiplicative spatial models of supply and demand: Theory and sensitivity analysis. *arXiv preprint arXiv:1101.3719*.
39. Massachusetts Department of Transportation (2012) 2010/2011 massachusetts travel survey. [Online; accessed 17-March-2016].
40. CTPS (2013) Methodology and assumptions of central transportation planning staff regional travel demand modeling.

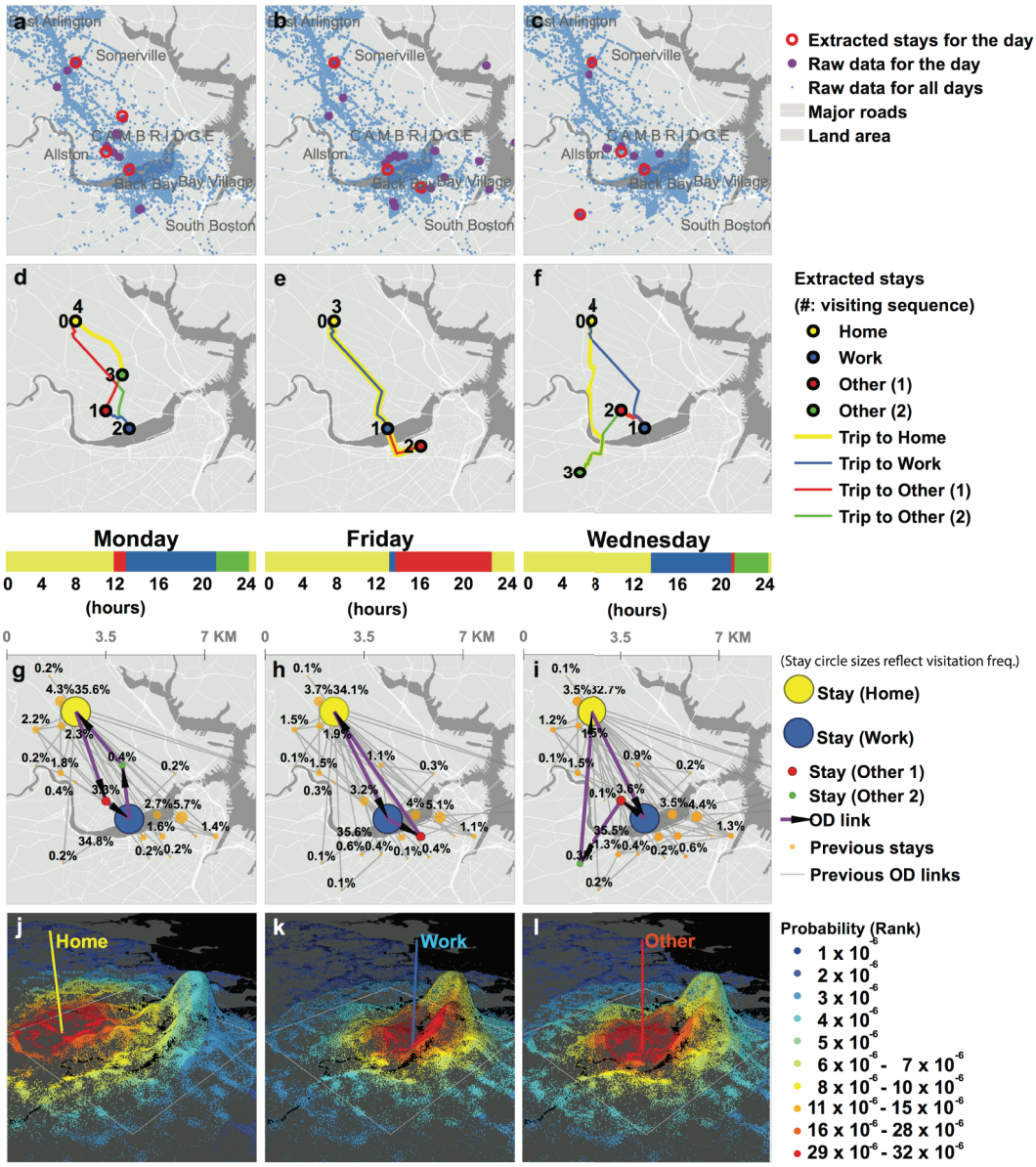

Fig. 1. Extraction of stays and daily journeys from raw cell phone data. (a-c) Stay locations extracted from the self-collected cell phone records of a student in three sample days. (d-f) Illustration of trips between consecutive stays in each day. (g-i) Visitation frequency of all locations, counting from the first day of the observation period to the current day. For this individual, *home* and *work* stays dominates all the visits. Highlighted arrows mark the trips on that day. The time-bar above each sub-figure is color-coded by activity type based on each stay's duration. (j-l) Illustration of the rank-based EPR model. To illustrate different cases we use the individual's *home, work*, and one *other* location as trip origins. The potential trip destinations are color-coded by different chosen probabilities based on their rank. The closer a location is to the origin, the higher the probability it has to be chosen. The height of the dots represents the density of destinations in the surrounding region. The most dense place for *other* type of activities is in downtown Boston.

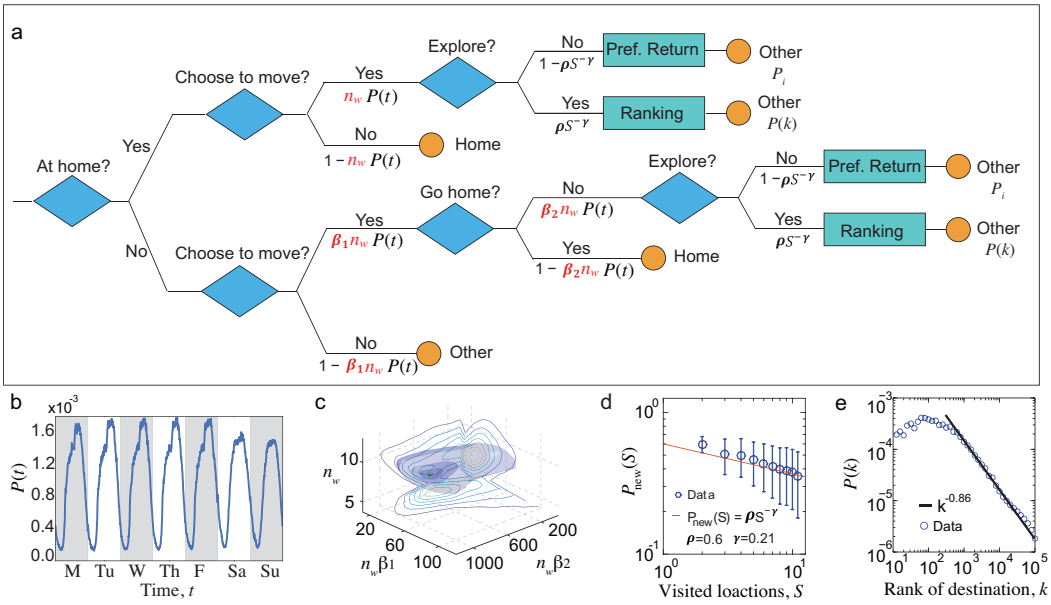

Fig. 2. Flow chart of TimeGeo and input features extracted from active CDR users. (a) Spatial and temporal choices per time step. Three individual specific parameters control temporal patterns, including the *weekly home-based tours* (n_w), *dwell rate* (β_1), and the *burst rate* (β_2). n_w influences the travel likelihood when a person is at home, $\beta_1 n_w$ influences the travel likelihood when a person is out of home, while $\beta_2 n_w$ influences the likelihood of performing consecutive out-of-home activities. (b) $P(t)$ shown here is the empirical travel circadian rhythm in an average week measured from data for active non-commuters (who have no journey-to-work trips). (c) Joint distribution of $\beta_1 n_w$, $\beta_2 n_w$ and n_w for active non-commuters in the CDR data set. The two-dimensional marginal distributions are shown by the contour plots. The green dot is the most probable parameter value combination with $n_w = 6.1$, $\beta_1 n_w = 22.4$, $\beta_2 n_w = 508.0$. (d) Empirical probability to visit a new location P_{new} as a function of distinct visited locations S, it follows $P_{new} = 0.6S^{-0.21}$. (e) Empirical probability of choosing the rank k location as a trip destination follows $P(k) \sim k^{-0.86}$.

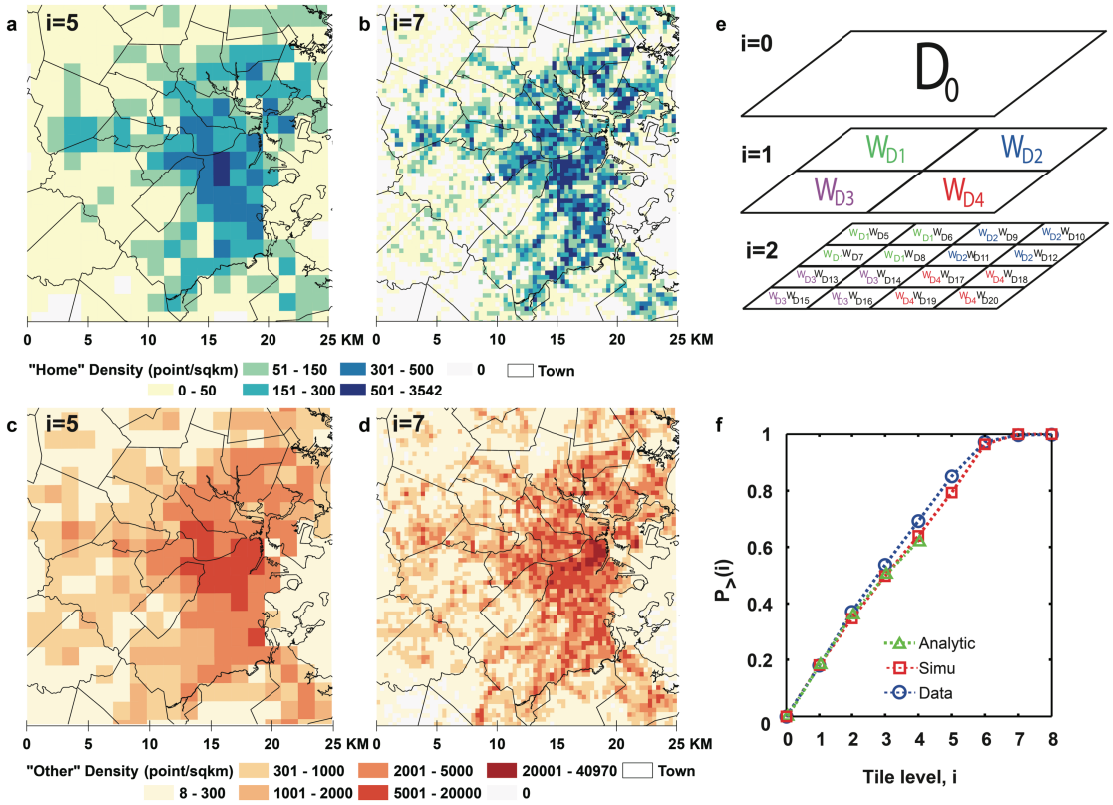

Fig. 3. **The multiplicative cascade analysis framework.** (a,b) The distribution of *home* locations in the Boston area at two different resolutions. (c,d) The distribution of *other* locations at two different resolutions. The variance of both distributions and their correlations depend on the resolution of the grids, or the cascade level i. At the scale with larger grid-cells, the number of non-residential (*other*) locations has higher correlation with the distribution of *home* locations; while at the scale with smaller grid-cells separation between residential and *other* land-use types are observed. (e) Illustration of the hierarchical cascade process generating trip demand D. Each tile is repetitively divided into 4 smaller tiles. The density of locations in each tile is controlled by the cascade generator W at each tile level. (f) $P_>(i)$ is the probability of an exploration trip going outside their origin tiles at level i at 8 tile levels with tile side-length from 24km to 187m (the entire Boston Metro Area, larger than the area shown in the maps, is set as a 48km square). Results show the calculation with the multiplicative cascade framework, in the simulation and measured by the mobile phone data.

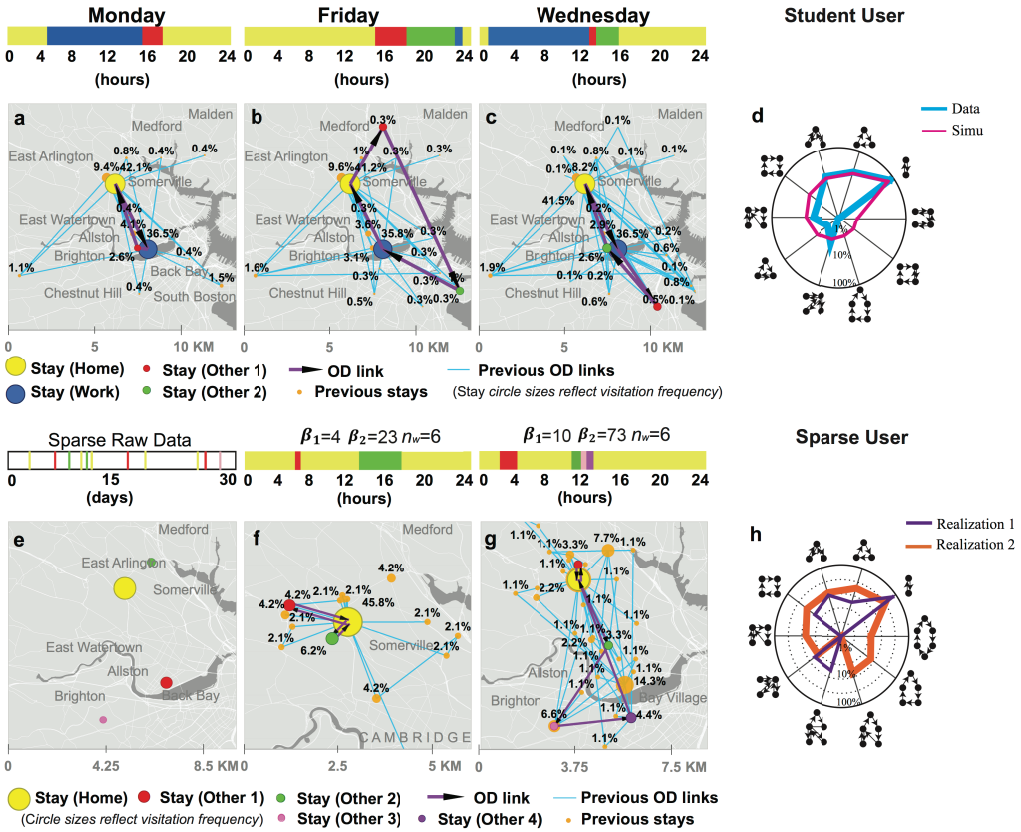

Fig. 4. Simulation of daily trajectories of one active commuter and one sparse user. (a-c) Simulated trajectories of the student with self-collected cell phone records. Three sample days are shown here. The trips for each sample day are in purple, and the visitation frequency of each location are calculated until the sample day and represented by the circle sizes. (d) Distributions of daily mobility motifs for the active commuter's data vs. simulation. The model captures well the higher propensity of motifs with node-sizes 2 and 3 as well as some other occurrences. (e) A sample sparse user with 10 stays at 4 distinct locations in an observation-period of 30 days. (f-g) Two different realizations for simulating the same sparse user with different parameter values. The first realization uses $n_w = 6$, $\beta_1 = 4$, $\beta_2 = 23$. The second realization uses $n_w = 6$, $\beta_1 = 10$, $\beta_2 = 73$. Larger values of β_1 and β_2 generate more consecutive out-of-home activities and more daily visited locations. (h) Distributions of daily mobility motifs for the two realizations of the same sparse user using different parameter values. With small n_w, $\beta_1 n_w$, and $\beta_2 n_w$ values the person is likely to have simple motifs, while large parameter values lead to more complex daily activity chains.

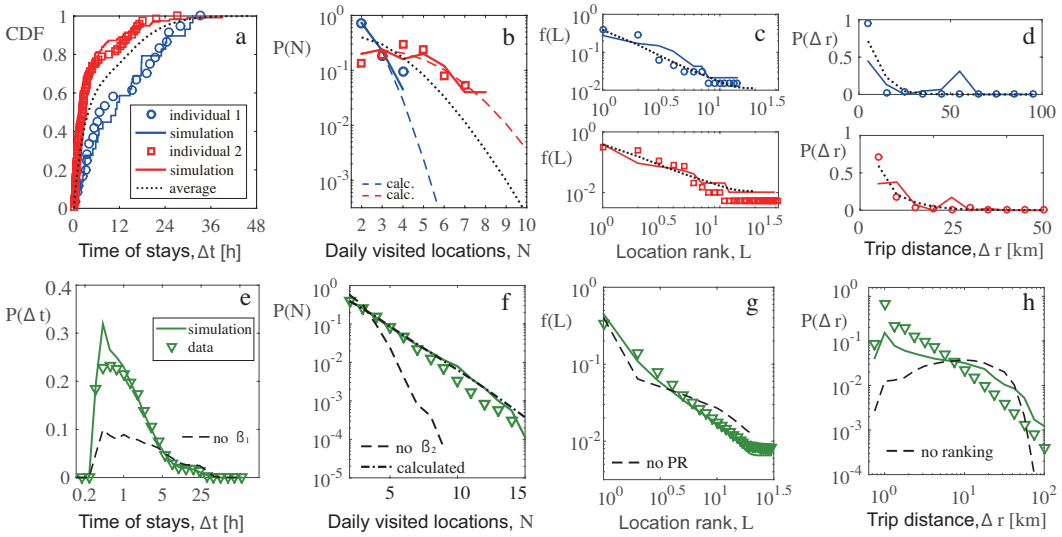

Fig. 5. Mobility patterns for different individuals and population distributions for non-commuters. The top panels (a-d) present comparison of mobility patterns for three representative non-commuters. Individuals 1 and 2 represent two extreme cases, one has shorter stays (shown in squares, $n_w = 10.86$, $\beta_1 = 6$, $\beta_2 = 41$) and the other travels less frequently (shown in circles, $n_w = 5.51$, $\beta_1 = 1$, $\beta_2 = 36$). The third case represents an average non-commuter and is simulated using the median parameter values of n_w, β_1 and β_2. (a) Stay duration distribution. (b) Number of daily visited locations. The Markov modeling framework allows the calculations of the number of visited locations per day, as is shown as dashed lines, and is discussed in SI Appendix. (c) Ranked frequency of locations with visitation rank L_i. (d) Distribution of trip distance. The bottom panels (e-h) show aggregated mobility patterns for all active non-commuters. (e) Activity duration distribution $P(\Delta t)$. The model without differentiating *home* and *other* sates (setting $\beta_1 = 1$) is considered as a benchmark here. In this case, stays with short duration are underestimated. (f) The distribution of the number of daily visited location $P(N)$. Both the model's calculation and simulation results are shown. It shows the need for the β_2 parameter in the model. (g) Visitation frequency $f(L)$ to the L^{th} most visited location follows the form $f(L) \sim L^{-1.2\pm0.1}$. The benchmark shows the result without the preferential return mechanism. (h) Trip distance distribution $P(\Delta r)$ extracted from data, and simulation results using a r-EPR mechanism, compared with the random selection of exploration locations (not using the rank-based selection mechanism).