Supporting Information

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 July 3, 2016

1 Mobile Phone Data Preprocessing

In the anonymous call detail record (CDR) data set that we obtained for the study area in 2010, each record is in the following format: [anonymized user ID, longitude, latitude, time stamp (in seconds)] . The coordinates of the recorded locations are estimated by the data provider using standard triangulation algorithms. The location information is in much higher resolution (with an accuracy of 200 to 300 meters) compared to that in the traditional tower-based CDR data [1–3]. This finer granularity enables us to identify user location more accurately and allows us to apply data mining methods previously tested for GPS records [4–6].

1.1 Extracting stays. The first step in the data processing pipeline is to identify users' stays and pass-bys (which are records made when users are conducting activities or traveling) from the raw data. As illustrated in Fig. S1, a *stay-point* is identified from a sequence of consecutive mobile phone records based on spatial and temporal thresholds. The spatial threshold is a roaming distance when a user is staying at a location, related to the accuracy of the underlying location positioning technology. In this study, we set the roaming distance of a stay-point as 300 meters. The temporal threshold is the minimum stay time (e.g., 10 minutes) at a location, measured as the duration between the first and the last record observed at a stay-point. Once a stay-point is identified, its location is set as the centroid of all records belonging to that stay-point (e.g., $s1$ in Fig. S1 is the centroid of mobile records $p3$, $p4$, and $p5$).

The second step is to cluster stay-points into stay-regions, since stay-points identified from a user's different trajectories over time may refer to the same location although the triangulated coordinates may not be exactly the same. We use a grid-based clustering method to cluster stay-points into stay-regions. The advantage of the grid-based clustering method is that it sets the output cluster sizes—which is desirable when we know that each location has a bounded size and the accuracy of the records is within a threshold. In this study, the maximum stay-region size is set as $d = 300$ meters. The procedure to perform grid-based clustering is as follows: First, divide the entire region into rectangular cells of size $d/3$. Next, map all the stay-points to each cell. Then, iteratively merge the unlabeled cell with the maximum stay-points and its unlabeled neighbors to a new stay-region. Once a cell is assigned to a stay-region, it's marked as labeled. [5] discusses details of this method. Illustrated in Fig. S1, the three stay-points $s1$, $s2$, and $s3$ are clustered into one stay-region $r1$.

Among the 1.92 million users of the raw CDR data obtained for the Greater Boston region—bigger than *Metro Boston*

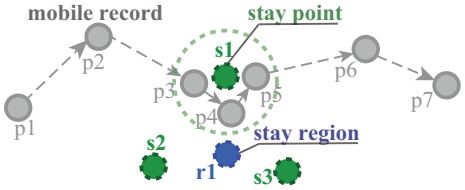

Fig. S1. Extraction of stays. Gray dots are raw mobile phone observations. Green dots are extracted stay-points, and blue point represents the stay-region.

(which includes 164 cities and towns) defined by the Boston metropolitan planning organization (MPO) [7]—1.66 million users (86%) have at least one stay observation longer than 10 minutes in her raw CDR data over the 6-week period).

For the self-collected cell phone traces provided by a volunteered individual for this research, different from the CDR data, a record is made when the mobile application detects a significant spatial movement. It is more likely that only one record are kept in the data set at a stay location. To cope with this characteristics, we (1) extract stay-points and stay-regions with spatial and temporal thresholds; (2) go through the records to search for points that are close (e.g. within 200 meter) to the existing detected stay-regions; and (3) add the record as a stay-region (given the data recording mechanism of the mobile application.)

1.2 Identifying location types: *home*, *work*, and *other*. To analyze and model urban mobility, we need to identify users' visited location types. For each extracted stay-region, we categorize it as *home*, *work*, or *other*. We label the most frequently visited stay-region during weekday nights (between 7pm of first day and 8am of second day) and weekend as the *home* stay-region. 1.44 million users (75% of the 1.92 million) are identified with home.

For a non-home stay, if its start time is during weekday daytime (between 8am and 7pm), it is defined as a *potential work stay*. The assumption to label a potential work stay into a *work stay* is based on the rationale and historical evidence [8, 9] that for a given visitation frequency, trips with longer distance are more likely to be work trips than those with shorter distance, which are more likely to be for non-work purposes (e.g., grocery shopping near home). For a user who has an identified home location, we label her potential work stay-region i as a *work* place if its distance from home (d_i') times its visitation frequency (n_i') is the maximum among all potential work stay-regions. We also restrict that a user's visitation to her *work* stay-region should not be less than 3

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

times in the observation period (i.e., $n_i' >= 3$), and its distance from home should not be too short (e.g., d_i'>500 meter). When a potential work stay-region fulfills the above criteria, it is labeled as *work*. Otherwise, it will be labeled as *other*. For validation purpose, we focus on users whose home is within the Metro Boston area defined by the Boston MPO [7]. Among the 1.44 million users with home, 0.78 million have home within Metro Boston, and 0.66 million have home outside Metro Boston. Among the 0.78 million users whose home are identified within Metro Boston, 0.42 million have been identified with work stay-regions. Table S1 summarizes the statistics for users with stays, homes and/or works derived from the CDR data set. If users have few records, it will be difficult to estimate their mobility parameters. We filter users who have more than 50 identified stays and at least 10 home stays in the observation period as *active* users, and derive a set of 0.177 million such users in the study area.

Table S1. Summary of CDR user statistics

	Millions	Percentage
Users in the raw data set	1.92	100%
Users with stays ($duration \geqslant 10min.$)	1.66	86%
Users with "home"	1.44	75%
within Metro Boston	0.78	41%
outside Metro Boston	0.66	34%

1.3 Validating *home* and *work*. By using the 2010 census population data and the 2006-2010 Census Transportation Planning Products (CTPP) data, we validate the identified home and work locations of the 0.78 million users within Metro Boston at the city and town level. To expand these 0.78 million users to population of the Metro Boston, the number of home stay-regions are aggregated to Census tracts of the Metro Boston. An expansion factor is calculated for each tract as the ratio of the 2010 Census population and the number of residents identified in the CDR data set. For Census tracts with fewer than 10 CDR residents (around 10 in the study area), the expansion factor is set to 0 to ensure that we do not overweight users that are not representative for a given Census tract.

Fig. S2 shows respectively the comparison of (a) residential and (b) employment population at town-level between 2010 Census data and the CDR estimates, and between the 2006-2010 Census Transportation Planning Products (CTPP) [10] data and the CDR estimates, for both before and after expanding the CDR data. The town-level correlation between the CTPP employment data and the estimated CDR employment is 0.99, and the sample expansion method adjusts well for the difference in magnitude. The total expanded CDR users with workplace is 2.3 million, while the CTPP reports a total of 2.1 million. This strong correlation is noteworthy, considering that each user's home and work locations were expanded based on their home location only.

2 Modeling Commuters

Commuters have three types of location including *home*, *work*, and *other*. We model *work* as an activity occurring at a fixed location (determined from the CDR data) with predetermined fixed duration.

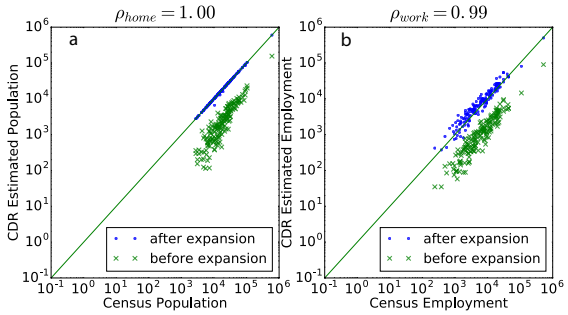

Fig. S2. Validation of the *home* and *work* labeling for the 0.78 million CDR users with detected home in the Metro Boston at the city and town level. (a) 2010 Census population v.s. CDR estimated residents at town-level before and after population expansion. (b) 2006-2010 CTPP [10] workers v.s. CDR estimated workers at town-level before and after population expansion.

2.1 CDR data and temporal features of the *work* activity. From common knowledge, we know that the majority commuters usually go to work in the morning and finish work in the late afternoon, and they may have work breaks during work hours. However, we find that the CDR data do not seem to capture well the temporal features of the work activities—namely, the work start time, duration and the presence of work-breaks. Fig. S3 compares the distribution of the start time and duration for work activity between CDR active users and the 2009 National Household Travel Survey (NHTS). It can be seen from the NHTS that two peaks (around 8am and 12pm, respectively) exist in the work start time and two peaks (around 4 hours and 8 hours, respectively) in the work duration. The two peaks in start time and duration are caused by breaks during working hours. In fact, both the 2009 NHTS and the 2010 American Time User Survey (ATUS) [11] show that around 20% workers have 1 work-break outside their workplace (which generates trips from workplace during the break.)

To overcome the shortcoming of the CDR data, we model the fixed work time for commuters detailed in the following subsection. If large scale mobile phone data with high frequency are available in the future, work time could be observed from mobile phone data.

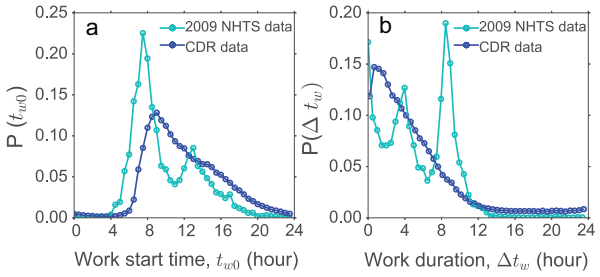

Fig. S3. Distribution of observed work start time and duration Marginal distribution of (a) work start time and (b) work duration. Note: The CDR data does not compare well with the survey data, not capturing well the beginning and end of the work trips.

2.2 The modeling approach. To model the fixed work activity for a commuter, we sample from a joint distribution of work start time and duration, and introduce stocastically a work-

break to allow for flexible activities during the break period. To be more specific, the detailed steps are as follows:

- We generate a pair of work start time (t_{w0}) and work duration (Δt_w) from a 2-D Gaussian mixture distribution for each commuter (see Section 2.3 for details). Note that this t_{w0} is simply the beginning of work in a day (on a weekday), and Δt_w simply measures work duration from beginning to end in the day.

- We introduce a parameter (θ) to stocastically determine if a commuter will take a work-break. We randomly generate a number $x \sim U(0,1)$. If $x<\theta$, the commuter will have a work-break outside workplace. We fix the proportion of commuters to have work-break outside workplace, $\theta = 0.20$.

- If the commuter takes a work-break, we then generate a work-break duration (Δt_b) and a work-break start time t_{b0} from two distributions, respectively (see Section 2.4 for details). With the generated t_{b0} and Δt_b, we then update the work time-slots which are now split by the work-break.

On a weekday, for a commuter with predetermined t_{w0} and Δt_w, and t_{b0} and Δt_b (if she takes a work-break), the TiemGeo model is applied to fill the rest time slots that are not occupied by work activities. Fig.S4 demonstrates how the model works for commuters.

As illustrated in Fig. S4 (a), on a sample weekday, a user is predetermined to stay at workplace from 9 am to 5 pm with no work-break. The proposed TimeGeo model will fill the rest time slots. The user will make a trip to work at 9 am, independent of her activity before 9 am (the blue slot). She will move from work after 5 pm (the green slot), but whether to go *home* or to *other* location is simulated by the TimeGeo model. Similarly, in Fig. S4 (b), a user is first predetermined to work from 9 am to 5 pm and take a break outside workplace between 1 pm and 2 pm. The model works the same as in (a) for the time slots before 9 am and after 5 pm. For the work-break, the user will definitely make a trip from work after 1 pm (the cyan slot) and then move back to work at 2 pm. She decides where to go in the break based on the TimeGeo model. The user could visit multiple *other* locations during the break, simulated by the TimeGeo model.

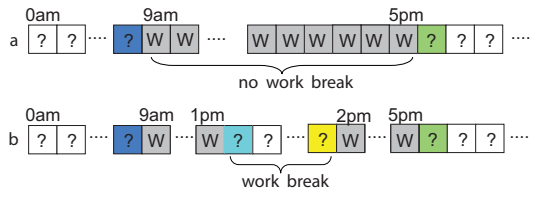

Fig. S4. Modeling commuters. For commuters with fixed work activity and (a) with no work-break, (b) with work-break. Work start time and duration, and break start time and duration are predetermined from distributions discussed later.

2.3 Work start time and duration. To characterize the statistic of work start time (t_{w0}) and work duration (Δt_w) from alternative data sources, we use NHTS and ATUS to estimate the joint distributions of t_{w0} and Δt_w. We fit the data with a mixture of multivariate normal distributions, and use a Gaussian Mixture Model (GMM) [12] to estimate parameters of the following joint distribution $f(\mathrm{x}|\mu_1,\Sigma_1,...,\mu_K,\Sigma_K) = \sum_{k=1}^{K}\pi_k \frac{1}{2\pi\sqrt{|\Sigma_k|}}\exp\left(-\frac{1}{2}(\mathrm{x}-\mu_k)^\top\Sigma_k^{-1}(\mathrm{x}-\mu_k)\right)$, where $K(>0)$ is the number of modes (distinct local maxima), $\mathrm{x} = (A,B)^\top$, $\mu_k = (\mu_{kA},\mu_{kB})^\top$, $\Sigma_k = \begin{pmatrix} \sigma_{kAA}^2 & \mathrm{cov}_{kAB} \\ \mathrm{cov}_{kAB} & \sigma_{kBB}^2 \end{pmatrix}$, A stands for t_{w0}, B stands for Δt_w, π_k is the mixing coefficient, $\sum_{k=1}^{K}\pi_k = 1$.

We find that three clusters ($K=3$) best fit the empirical data. It is in good agreement with existing studies on the temporal behavior of workers' daily activity patterns found in [13] (e.g., early workers, regular workers, and late workers). Fig. S5 presents the marginal distributions of t_{w0} and Δt_w estimated from the 2010 NHTS and 2009 ATUS data, respectively. The results are quite similar, and we use the set of parameters from NHTS to jointly generate t_{w0} and Δt_w: $K=3$, $\pi_1=0.17$, $\pi_2=0.29$, $\pi_3=0.53$, $\mu_{1A}=12.8$, $\mu_{2A}=7.9$, $\mu_{3A}=7.6$, $\mu_{1B}=6.6$, $\mu_{2B}=7.5$, $\mu_{3B}=9.0$, $\sigma_{1A}=3.7$, $\sigma_{2A}=1.5$, $\sigma_{3A}=1.0$, $\sigma_{1B}=4.4$, $\sigma_{2B}=3.2$, $\sigma_{3B}=0.9$, $\mathrm{cov}_{1AB}=-4.3$, $\mathrm{cov}_{2AB}=-2.6$, $\mathrm{cov}_{3AB}=-0.3$. (Time units are in hours.)

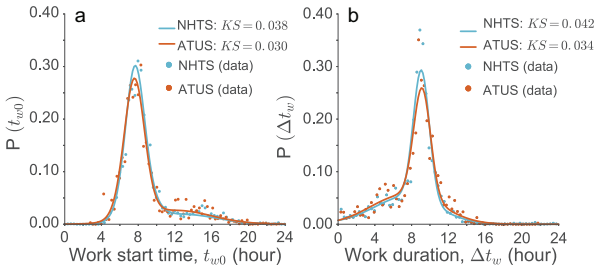

Fig. S5. Marginal distribution of work start time (t_{w0}) and duration (Δt_w). Note: t_{w0} represents the beginning of work in a day. Δt_w simply measures work duration from beginning to end in a day. Parameters are estimated from a 2-dimensional GMM.

2.4 Work-break start time and duration. To characterize work-breaks from data, we first estimate the distribution of their duration (Δt_b) from the NHTS and AUTS, shown in Fig. S6 (a). The probability density of Δt_b follows a log-normal distribution, i.e., $P(\Delta t_b) = \frac{1}{\sqrt{2\pi}\sigma\Delta t_b}e^{-\frac{(\ln\Delta t_b-\mu)^2}{2\sigma^2}}$, where $\mu = 3.9$, $\sigma = 0.9$ (time unit in minute). Note, in the simulation, when generating Δt_b, we make sure that $\Delta t_b < \Delta t_w$.

From common knowledge, work-break start time (t_{b0}) peaks in the middle of work, although it may occur anytime during work. We measure the distribution of the normalized deviation of work-break midpoint (t_{bm}) from the work midpoint (t_{wm}), noted as $D_{bw} = \frac{t_{bm}-t_{wm}}{\Delta t_w - \Delta t_b}$, from the 2009 NHTS data, shown in Fig. S6(b). We find that D_{bw} follows the following truncated Cauchy distribution $P(D_{bw}) = \frac{1}{\pi\gamma}\cdot\frac{\gamma^2}{(D_{bw}-x_0)^2+\gamma^2}$, where $x_0 = 0.0$, $\gamma = 0.1$, $-0.5 < D_{bw} < 0.5$ (since a work-break has to start after work starts and end before work ends). For the simulation of work-breaks, we randomly draw a D_{bw} from the truncated Cauchy distribution, and then determine the work-break start time according to $t_{b0} = t_{bm} - 0.5\Delta t_b = t_{w0} + (\Delta t_w - \Delta t_b)(0.5 + D_{bw})$. With the generated work-break start time and work-break duration, we then update the work time-sots, splitting them by the work-break.

Although the above parameters are estimated using NHTS data for consistency check, we do not think that expensive

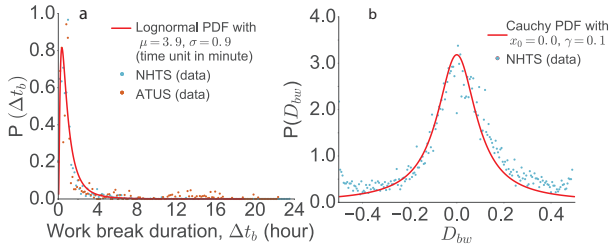

Fig. S6. Model for work-break. (a). Empirical distribution of work-break duration Δt_b. (b) Empirical distribution of the normalized deviation of work-break midpoint from the work midpoint, D_{bw}, $(-0.5 < D_{bw} < 0.5)$.

travel surveys are necessary for the purpose of modeling typical work activity (i.e., work start time, work duration, work-break duration, work-break start time).

3 Model Parameter Estimation

3.1 Population travel circadian rhythm. For each day in an average week, we measure the population travel circadian rhythm as the probability of traveling to and from flexible activities in every 10-minute time slot t, which is denoted as $P(t)$. Since commuters' work activities are modeled as fixed choices, the probability $P(t)$ for commuters does not include trips to and from the work activity. In other words, in a time slot t, only if a commuter travels to and from either *home* or *other* (not *work*), the trip is counted towards the probability of $P(t)$. We measure $P(t)$ separately for commuters and non-commuters, as shown in Fig. S7. For an average commuter, her travel rate to and from flexible activities is not high during working hours, but peaks around 6 pm on weekdays; the peaks at weekend are higher than those on weekdays. For an average non-commuter, her travel rate of flexible activities during weekday working-hours is higher than that of an average commuter; and the peaks of the travel rate are lower at weekend than on weekdays.

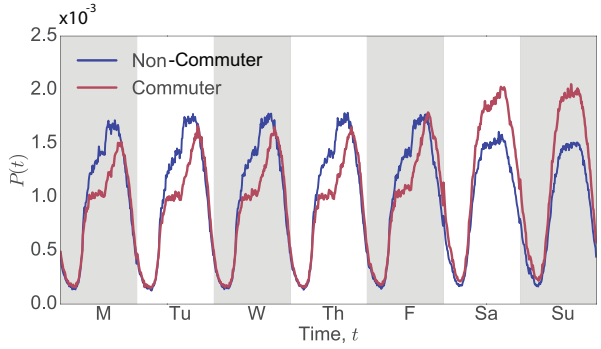

Fig. S7. Circadian travel rhythms for commuters and non commuters. Note: Because the fixed activity—*work*—is not determined by the Markov model, travels to and from *work* for commuters are excluded from the measure of $P(t)$.

3.2 Exploration and preferential return (EPR) parameters. In this section, we illustrate that while we estimate global parameters from the CDR data to simulate the EPR mechanisms, in contrast to the results of the original EPR model, we find differences among individuals. The individual heterogeneity are now captured by the newly introduced weekly trips (n_w) and the two individual parameters of the Markov model. Fig. S8 (a) shows that for different S groups, the number of visited distinct locations S(t) versus time follows $S(t) \sim t^{\mu}$. For S group: 5-10, $\mu = 0.54$; S group: 10-20, $\mu = 0.68$; S group: 20-30, $\mu = 0.76$; S group: 30-40, $\mu = 0.80$.

Fig. S8 (b) shows that for users with different distinct locations (S), their visitation frequency to the Lth most visited locations f_L follows: $f_L \sim L^{-\xi}$, with $\xi = 1.2 \pm 0.1$, similar to the finding in [2]. We estimated global EPR parameters for model $P_{new} = \rho S^{-\gamma}$, with $\rho = 0.6$ and $\gamma = 0.21$. Our finding is consistent with [2], which showed that $\xi = 1 + \gamma$, if $\gamma > 0$.

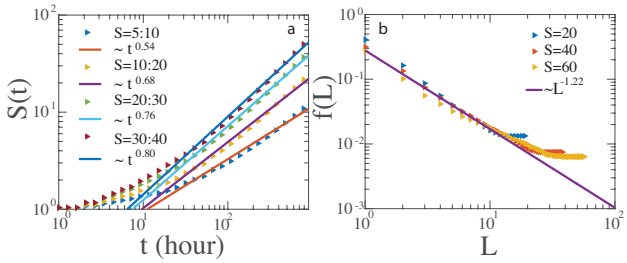

Fig. S8. Empirical results on the exploration and preferential return (EPR) parameters estimated from the CDR data of active users. (a) the number of visited distinct locations S(t) versus time for different S groups, (b) the visitation frequency to the Lth most visited locations f_L follows: $f_L \sim L^{-\xi}$, with $\xi = 1.2 \pm 0.1$

3.3 Individual mobility parameters. Fig. S9 illustrates the Markov model of transition probabilities at *Home* (H) or *Other* (O) in a 10-minute time-slot t for a non-commuter. We show in this example two *other* states to demonstrate the current *other* state, and a consecutive new *other* state. In time-slot t, when an individual is at *home*, her probability of staying home is $P_1 = 1 - n_w P(t)$. Her probability of traveling to an *other* location is $n_w P(t)$. When she is not at home, but at an *other* location, the individual is in an *active* state—her probability of staying at the current *other* location is $P_2 = 1 - \beta_1 n_w P(t)$, and her probability of visiting a consecutive *other* location is $P_3 = \beta_1 n_w P(t) \beta_2 n_w P(t)$. When she moves from an *other* state, she can either choose to go to an additional *other* location with probability P_3, or go *home* with a probability $P(O \to H) = 1 - P_2 - P_3 = \beta_1 n_w P(t)(1 - \beta_2 n_w P(t))$. To ensure that a person will go home at the end of the day, we add a condition that after certain hour in the late afternoon (e.g., 5 pm) the individual's returning home probability is the maximum value of $P(O \to H)$ and $(1 - \frac{P(t)}{max(P(t))})$.

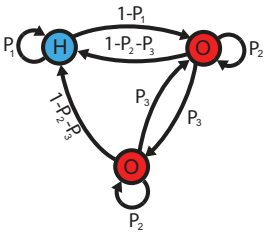

Fig. S9. Illustration of the Markov model of transition probabilities between *Home* and *Other* state in a 10-minute time-slot t for a non-commuter.

We now show the empirical individual parameters measured for active commuters (133, 448 individuals) and non-commuters (43, 606 individuals). The joint distribution of parameters n_w, $n_w\beta_1$ and $n_w\beta_2$ is in Fig. S10. The median values of n_w, $n_w\beta_1$ and $n_w\beta_2$ for non commuters are 7.4, 34.2, and 355.6, while the values for commuters are 5.7, 21.2, and 286.7. The two dimensional marginal distributions are shown by the contour plots.

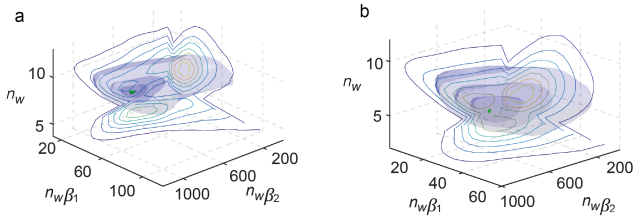

Fig. S10. Joint distribution of parameters n_w, $n_w\beta_1$ and $n_w\beta_2$. Parameter distribution for (a) non-commuters, and (b) commuters.

The marginal distribution of each parameter is in Fig. S11. The distribution of n_w and $n_w\beta_1$ could be approximated by log-normal distributions while the distribution of $n_w\beta_2$ is approximated by Weibull distribution. The corresponding estimation results are also shown in Fig. S11.

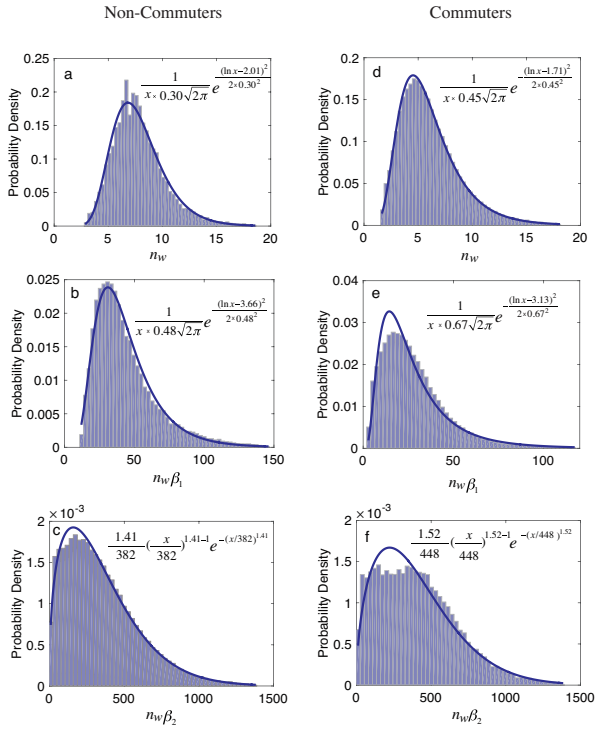

Fig. S11. Marginal distribution of parameters n_w, $n_w\beta_1$ and $n_w\beta_2$. for (a-c) non-commuters, and (d-f) commuters.

4 Model Simulation and Validation

TimeGeo is a platform that can be used to simulate individual daily mobility trajectories at fine temporal and spatial resolution, and help researchers, planners, and policy makers understand urban mobility from individual level to metropolitan level.

4.1 At the individual level. With technology that can capture individual spatiotemporal mobility records more accurately and completely, such as the one demonstrated by the student volunteer's data recorded by a mobile phone application (Fig. 4), TimeGeo can be used to simulate individual daily mobility patterns similar to the observed ones shown in mobile phone data without additional data inputs. Fig. S12 presents the comparison between the mobile phone data for this volunteer user and simulated results on the distributions of (a) daily visited location numbers, (b) activity stay duration, (c) location visitation frequency for the *Lth* most visited locations, and (d) trip distance. (The user's trip distance is dominated by travel between home and school.)

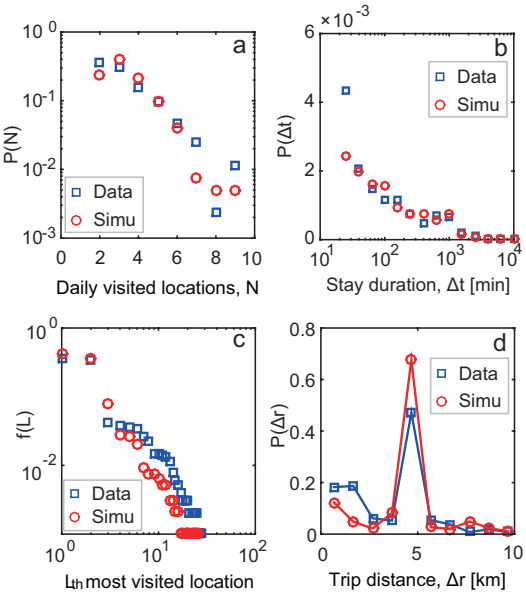

Fig. S12. Simulation results for the volunteer student (shown in Fig. 4): Distributions of (a) number of daily visited locations; (b) stay duration; (c) visitation frequency of the Lth most visited location; and (d) trip distance.

To give another example, Fig. S13 presents a comparison between observed and simulated trajectories of two representative individuals. The background red color shows the density of *other* locations. In the simulation, the home location and the number of visits to home are kept unchanged for each person. Visits to other locations are modeled using the rank-based EPR mechanism. The β_1 and β_2 parameters for each person are calibrated using their CDR data. The weight of each line represents the visitation frequency. The star symbol shows the home location of the individual. It is noteworthy that both the CDR data and the simulated trajectories show that if a person's home is far from areas with dense *other* locations, she tends to travel longer distance more frequently. On the other hand, if the person's home is in the city center (with dense *other* locations nearby), her tendency to travel far is lower.

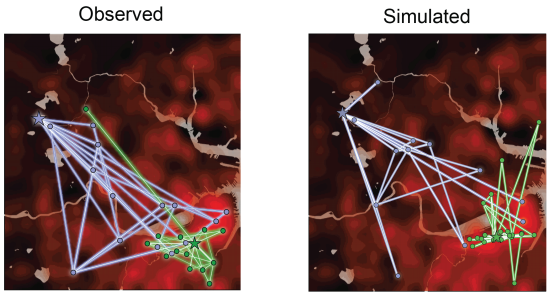

Fig. S13. Spatial comparison of observed vs. simulated trajectories for two representative individuals—one user in gray and the other in green.

4.2 Daily mobility motifs. In addition to the number of daily visited locations, the model is able to capture similar daily mobility motif distribution revealed from the CDR data. Fig. S14 (a) is the aggregated motif distributions for all active users. The empirical data are in blue and the simulation results are in green. As a guide to the eye, two dashed lines at 1% and 5% are shown in the figure. The most popular trip motif is traveling between two locations. Fig. S14 (b) shows the motif distribution for commuters and non-commuters separately. To show the less popular motifs clearer, we plot the distribution in log scale in the inset of each figure. In general, the more complex a motif is, the lower the percentage is.

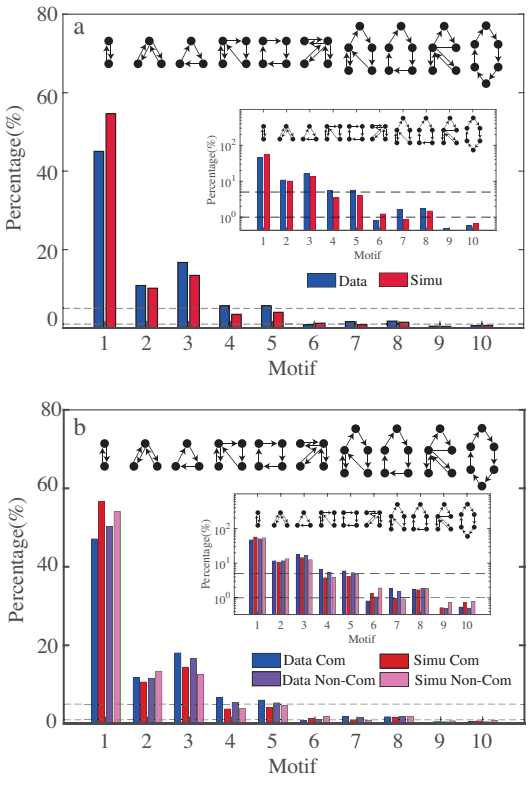

Fig. S14. Motif distributions. (a) Motif distribution for all active users, shown in linear and log scale (inset figure). The two dashed horizontal lines are respectively 1% and 5%. (b) Motif distribution for commuters and non commuters (inset figure is in log scale).

4.3 At the metropolitan level. With the assumption that young children are not represented by the CDR data set, we simulate the mobility trajectories in Metro Boston for population aged 16 years and over (to be consistent with the census data, e.g., American Community Survey). We expand the active phone users to the population aged 16 and over (i.e., 3.54 million people) in Metro Boston. We derive two sets of expansion factors, to expand active commuters to 2.10 million workers and expand active non-commuters to the rest 1.44 million non-workers, at the census tract level respectively (data available from American Community Survey). Fig.S15 shows the distribution of expansion factors for (a) commuters, and (b) non-commuters.

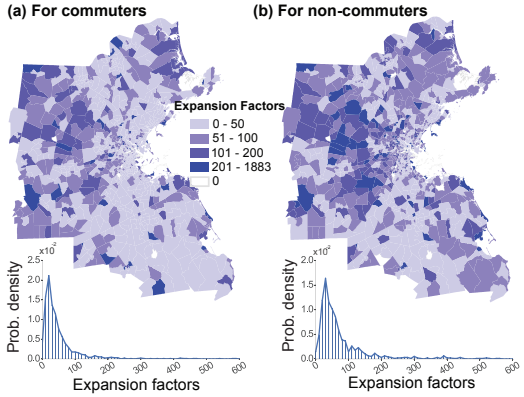

Fig. S15. Expansion factor distributions. Top figures show the spatial distribution of expansion factors to expand the active CDR users whose home are in the census tracts: (a) to expand commuters to the total employment population, and (b) to expand non-commuters to non-employment population (above 16 years old) in the census tract. Bottom figures show the probability density distributions of the expansion factors for (a) commuters and (b) non-commuters.

Fig.S16 shows the distribution on (a) stay duration (Δt), (b) daily visited location (N), (c) trip distance (Δr) for simulation of active phone users (for both commuters and non-commuters) as well as survey data, including the 2009 NHTS [14], and 2010-2011 Massachusetts Travel Survey (MTS) [15]. We also included sensitivity analysis for the simulation on the sample sizes and number of simulation days to demonstrate their implications on the simulation results. Note that we do not include the travel distance distribution from NHTS for comparison, given that spatial aspects of travel are affected more directly (than the temporal aspects) by urban form, which varies across the nation [16].

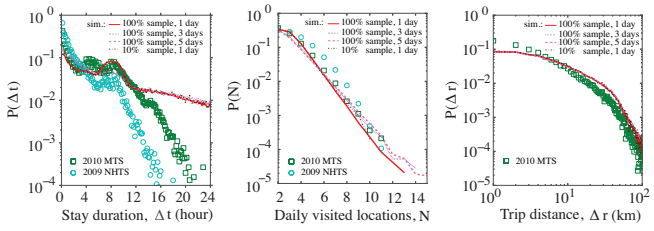

Fig. S16. Population distribution comparison: simulation and travel survey data—including 2009 National Household Travel Survey (NHTS), and 2010-2011 Massachusetts Travel Survey (MTS). (a) Stay duration distribution. (b) Daily visited location distribution. (c) Trip distance distribution.

The sensitivity analysis shows that when the simulation sample size is comparable to that of the survey, the number of daily visited locations of the simulation is similar to the survey. When the simulation sample size is larger than that of the survey, or the simulation days are more than 1 day (while the survey is only for 1 day), the simulation reveals higher proportion of large number of daily visited locations.

Based on the model discussed in the paper, we keep home and work locations and stay location records of active users to simulate daily trajectories of the expanded 3.54 million individuals who are over 16 years old and reside in the Metro Boston area. Note that our simulation allows trips to and from *other* type locations beyond the Metro Boston boundary as presented in the active mobile phone users' records. Fig.S17 compares the simulated daily trips per person by trip purpose and by time period with the Boston MPO travel demand model for years 2010 [7] and 2007 [17]. The MPO models follow the traditional four-step modeling of trip generation, trip distribution, mode choice, and trip assignment. The trip purposes include (1) home-based work (HBW), (2) home-based other (HBO), and (3) non-home-based (NHB). The time periods include AM peak (6 am-9 am), midday (MD) (9 am-3 pm), PM peak (3 pm-6 pm), and rest-of-day (RD) (6 pm-6 am). We also include the 2010-2011 MTS data for comparison.

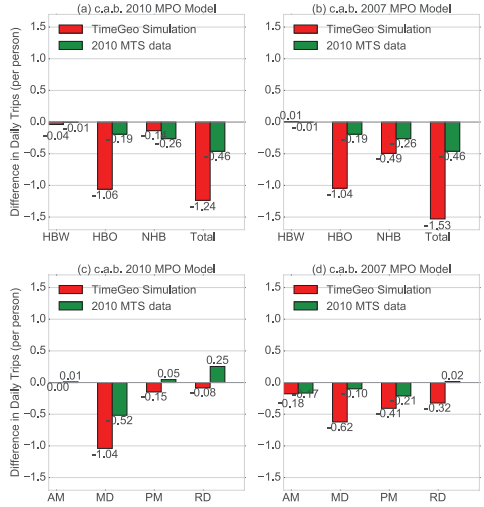

Fig. S17. Comparison against baseline (c.a.b.)—2010 and 2007 Boston MPO travel demand models—at the person level. Note: Values closer to zero mean simulation results are with small differences from the base line. Simulation is only for population older than 16 years old (3.54 million persons in 2010). Boston MPO 2010 model is for total population (4.46 million persons in 2010) excluding school trips (categorized as the HBO trips). Boston MPO 2007 model is for total population (in 2007) including all trip types. The 2010 MHT data presented here only include trips made by individuals aged 16 and over.

The comparisons between our simulation and the MPO models shown in Fig.S17 only include trips within the Metro Boston area, even though the TimeGeo model also simulated trips outside the region. Note that the Boston MPO 2010 model is for all aged population including 4.46 million persons (in year 2010) but excluding school trips, and the MPO 2007 model is for all aged population in year 2007 including all trip types (school trips are included in the HBO category). To make the comparison meaningful, we look at daily trips per person. In general, the simulated HBW trips, and trips in the AM, PM and RD periods are in good agreement with the MPO 2010 model. The simulated HBO and midday trips are noticeably less than those estimated by the MPO 2010 model. When considering all trips (including those beyond the metropolitan boundary) by TimeGeo, the results simulated by TimeGeo, the 2010 MPO model, the 2007 MPO model, and the 2010 MTS, for [HBW, HBO, NHB, daily] trips, are as follows: [0.74, 0.74, 1.42, 2.90], [0.67, 1.60, 0.63, 2.90], [0.63, 1.58, 0.99, 3.20], [0.62, 1.39, 0.73, 2.74]; for [AM, MD, PM, RD] trips, the results are [0.34, 0.86, 0.64, 1.06], [0.33, 1.50, 0.57, 0.54], [0.51, 1.08, 0.82, 0.78], [0.34, 0.98, 0.61, 0.80]— all units are in daily trips per person.

Fig. S18 shows the departure time of travel by trip purpose. The comparisons are among the TimeGeo simulation, 2009 NHTS, and 2010-2011 MTS (extracted for residents within Metro Boston). The patterns for HBW, and all trip purposes are similar among the three sources. The simulation does not have a morning peak for HBO trips—the potential reason might be that we are not simulating school trips, while travel surveys include those trips. The simulation has a higher fraction of NHB trips in the early evening compared with those of the surveys, which may be due to the fact that people tend to omit their short out-of-home stops when reporting in the travel survey.

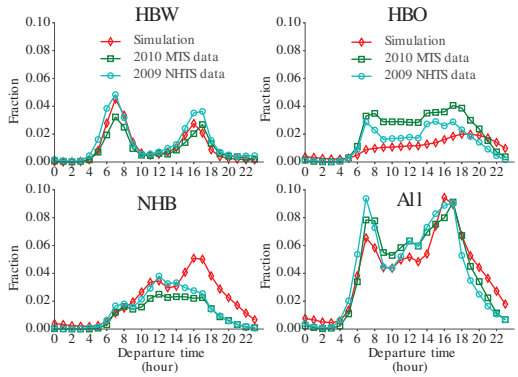

Fig. S18. Comparison of travel departure time among TimeGeo simulation, 2009 NHTS data, and 2010 MTS data, for trips by purpose of home-based work (HBW), home-based other (HBO), non-home-based (NHB), and all types (All).

Fig. S19 shows the comparison of origin-destination (OD) trips by trip purpose and by time period between the TimeGeo simulation and the Boston MPO 2010 model at the city and town level (with inter-town and intra-town trips separated). Correlation of the intra-town estimation between the two models are very good (the Pearson correlation coefficients are between 0.99 and 1.0), and those for the inter-town estimation are relatively lower. Since the employed population for the two sources are the same, the correlation between the simulation and the MPO model is very good for HBW and AM peak period. Due to the differences in population for non-workers (i.e., the TimeGeo simulation only includes population aged 16 years and above, while the MPO model includes population for all age groups), simulated trips for HBO and NHB purposes within the MPO region are systematically lower than the MPO model estimates. Meanwhile, since the TimeGeo simulation includes trips to or from *other* locations outside the MPO

boundary (as represented by the active mobile phone users' records), the simulated HBO and NHB trips can go beyond the Metro boundary. Therefore the simulated trips bounded within the MPO region can be less than the estimates of the MPO model. However, the correlation between the two models are still very high (e.g., above 0.75 even for inter-town trips).

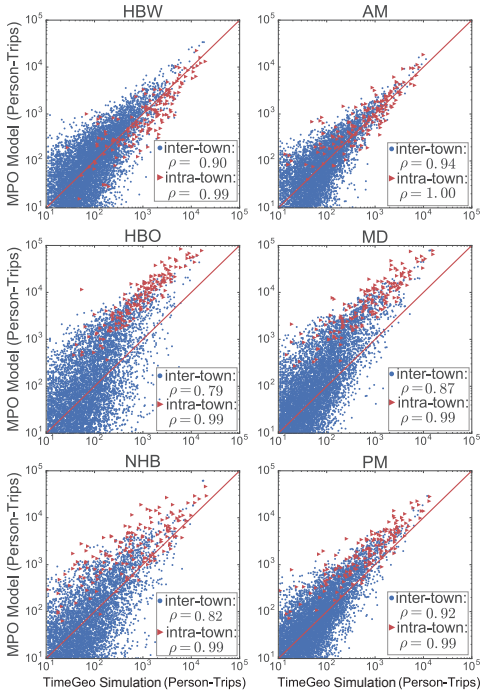

Fig. S19. Comparison of Origin-Destination(OD) trip tables at the city and town level: TimeGeo simulation (x-axis) and Boston MPO 2010 model (y-axis). Note: TimeGeo simulation is only for population who are aged 16 and over (i.e., 3.54 million persons) and whose trips are within the Boston MPO region, while the Boston MPO 2010 model is for population of all age groups (i.e., 4.46 million persons) within Metro Boston. The simulation allows trips to and from *other* type locations beyond the Metro boundary as presented in the active mobile phone users' records. The OD comparisons between the TimeGeo simulation and the MPO model shown here only include trips within the same region.

5 The Multiplicative Cascade Framework

5.1 Parameter estimation in β-log-normal cascade. The parameters of a β-log-normal cascade can be estimated from the moments of point density at different tile levels. A hallmark of multi-fractality is that the moments of the measure density $D'_i = D_i/|\Omega_i|$ are power functions of the resolution 2^i,

$$E[D'^q_i] \propto 2^{iK(q)} \qquad [S1]$$

where $K(q) = \log_2(E[W^q_D])$ is a concave function. Hence, multi-fractality holds if the log-log plots of the moments against resolution are linear and $K(q)$ is the slope of those linear plots. For β-log-normal cascades, $K(q)$ is a quadratic function,

$$K(q) = -(\log_2 P_D)(q-1) + \frac{V_D}{2}(q^2 - q) \qquad [S2]$$

with β parameter P_D and log-normal parameter $V_D = \sigma^2_{W_D}/\ln(2)$. There are different ways to estimate these parameters using the $K(q)$ function. A simple one is to use the empirical values of $K(0)$ and $K(2)$. Then $P_D = 2^{K(0)}$, $V_D = K(2) + K(0)$. If $K(0) = 0$, then $P_D = 1$ and the cascade is purely log-normal, and if $K(2) = -K(0)$, then $V_D = 0$ and the cascade is purely β. If the moments do not scale with resolution, one may use the local moment slopes at resolution level $i = 1, 2, ...$ to estimate $K_i(q)$ and obtain an approximation of resolution-dependent parameters P_{D_i} and V_{D_i}.

To estimate the correlation coefficient in the bivariate cascade generators $[W_{S_i}, W_{D_i}]$, we write their relationship with the measured density at tile i and $i-1$ as:

$$\begin{bmatrix} D'_i \\ S'_i \end{bmatrix} \overset{d}{=} \begin{bmatrix} W_{D_i} & 0 \\ 0 & W_{S_i} \end{bmatrix} \begin{bmatrix} D'_{i-1} \\ S'_{i-1} \end{bmatrix} \qquad [S3]$$

$$E[S'_i D'_i] = E[W_{S_i} W_{D_i}] E[S'_{i-1} D'_{i-1}] \qquad [S4]$$

$$E[W_{S_i} W_{D_i}] = \mu_{S'_i} \mu_{D'_i} + cov[W_{S_i}, W_{D_i}] \qquad [S5]$$

If $[W_{S_i}, W_{D_i}]$ are joint log-normal variables,

$$\sqrt{cov[W_{S_i}, W_{D_i}]} = e^{\sqrt{\rho_{LN_i} \sigma_{W_{D_i}} \sigma_{W_{S_i}}}} - 1 \qquad [S6]$$

ρ_{LN_i} is the correlation coefficient of $\ln(W_{S_i})$ and $\ln(W_{D_i})$. In the above equations, $E[S_i D_i]$ can be estimated from the density count at each i-tile, μ_{S_i}, μ_{D_i} are 1, $\sigma_{W_{D_i}}$ and $\sigma_{W_{S_i}}$ are estimated from the local slopes of the moment plot. Therefore ρ_{LN_i} could be estimated from the data. The calculation for β cascade is similar. The various moments of the empirical supply and demand density distributions $D'_i = D_i/|\Omega_i|$ and $S'_i = S_i/|\Omega_i|$ are shown in Fig. S20 (a-b).

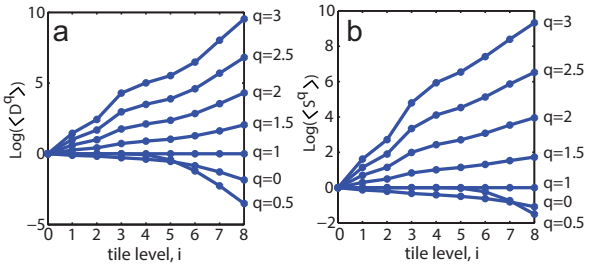

Fig. S20. Parameter estimation in the cascade point generation model. (a-b) Moments of the empirical density of demand D'_i and supply S'_i at different tile levels i and moments q. The slope of the moment plot is used to estimate the parameters of the β-log-normal cascade.

The slope of the moment 0 and 2 plot of home location $K_D(0)$, and $K_D(2)$ are in Table. S2. Thus we can calculate P_D and V_D accordingly. Similarly, we can use the moment plot of other location to infer P_S and V_S. $P = 1$ indicates pure log-normal cascade and $V < 0$ indicates the cascade is better represented by a pure β cascade. Thus for the home location distribution, at the 4 larger tile levels the cascade is approximately pure log-normal while at the 4 smaller levels the cascade is approximately pure β. The correlation coefficient between supply and demand could then be estimated. The result is also in Table. S2. It drops from 0.92 at the coarsest granularity to 0.23 at the finest granularity.

In the absence of data, these parameters can be used to simulate home and other location distributions in the area. The comparison of the real vs. simulated location distributions is shown in Fig. S21.

Table S2. Multiplicative cascade model estimation result

	Level							
	1	2	3	4	5	6	7	8
$K_D(0)$	0	0	0	-0.06	-0.37	-0.77	-1.06	-1.27
$K_D(2)$	0.61	0.38	0.75	0.35	0.26	0.47	0.7	0.77
P_D	1	1	1	0.95	0.77	0.59	0.48	0.42
V_D	0.61	0.38	0.76	0.28	-0.11	-0.3	-0.36	-0.49
P_S	1	1	1	1	0.97	0.87	0.71	0.59
V_S	0.69	0.45	0.85	0.44	0.24	0.17	-0.04	-0.35
ρ	0.92	0.59	0.72	0.76	0.77	0.78	0.45	0.23

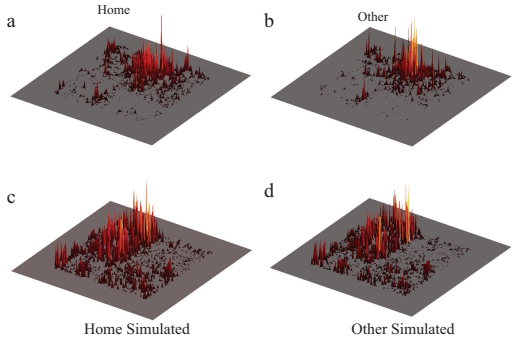

Fig. S21. Real and simulated location distributions. (a,b) are the density plots of the extracted home and other locations from the cell phone data. (c,d) are the simulated densities using the calibrated cascade point generation model.

The simulated distribution could reflect that there is a center of densely distributed locations in the Boston area, while in the peripheral areas the locations are sparsely distributed. The influence of the correlation coefficient and standard deviation in the log-normal cascade is shown in Fig. S22.

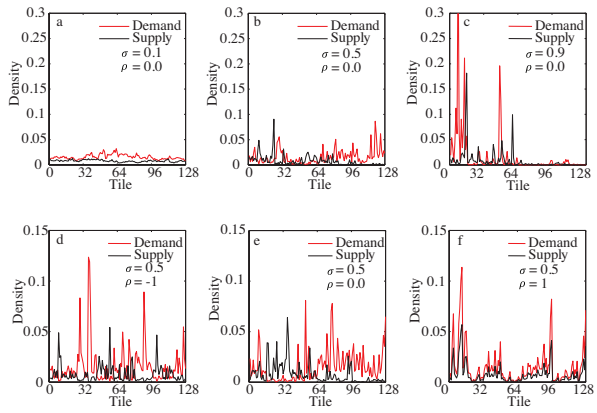

Fig. S22. Illustration of the joint log-normal cascade with different standard deviation and correlation coefficient.

5.2 Analytical characterization of trip distance. The multiplicative cascade framework can analytically characterize trip distances. It is important to predict distance traveled depending on the distribution of population (demand) and facilities (supply). $P_>(i)$ measures the probability that a generic trip goes outside its origin tile at resolution level i. To calculate $P_>(i)$, first of all, the probability to choose rank k location is

$$P(k) = \zeta k^{-\alpha} \quad [S7]$$

ζ is the normalization factor. The probability to choose within the closest k opportunities is:

$$P_<(k) = \zeta 1^{-\alpha} + \zeta 2^{-\alpha} + ... + \zeta k^{-\alpha} \approx \zeta \int_1^k x^{-\alpha} dx = \frac{\zeta}{1-\alpha}(k^{1-\alpha} - 1) \quad [S8]$$

Assume there are M opportunities in the region, then $P_<(M) = 1$

$$\frac{\zeta}{1-\alpha}(M^{1-\alpha} - 1) = 1 \quad [S9]$$

$$\zeta = \frac{1-\alpha}{M^{1-\alpha} - 1} \quad [S10]$$

To obtain tile exceedance probability $P_>(i)$, consider a generic trip with origin in Ω_0 and denote by $[S_{i,trip}, D_{i,trip}]$ the random supply and demand in the sub-region Ω_i where the trip originates. Then the probability to travel outside tile i is

$$P_>(i) = \int_1^M P_>(k) f_{S_{i,trip}}(k) dk \quad [S11]$$

where the probability density function (PDF) $f_{S_{i,trip}}$ can be calculated as follows. The PDF of $D_{i,trip}$ is related to the PDF of D_i as

$$f_{D_{i,trip}}(D) \propto f(i|D_i = D) \times f_{D_i}(D) \approx D f_{D_i}(D) \quad [S12]$$

and the conditional distribution of $[S_{i,trip}|D_{i,trip}]$ is the same as the conditional distribution of $[S_i|D_i]$. Therefore

$$f_{S_{i,trip}}(S) = \int_0^P f_{D_{i,trip}}(D) f_{S_i|D_i=D}(S) dD \quad [S13]$$

P is the population. To obtain the distribution of $\ln(S_{i,trip})$, we first calculate the PDF of $D_{i,trip}$. After some algebra, one obtains

$$f_{D_{i,trip}}(D) = \frac{1}{D\sqrt{2\pi}\sigma_D} e^{-[D-(m_D+\sigma_D^2)]^2/2\sigma_D^2} \quad [S14]$$

meaning that $\ln(D_{i,trip})$ has normal distribution with mean value $m_D + \sigma_D^2 = \ln(D_0 4^{-i}) + \frac{1}{2}\sigma_D^2$ and variance σ_D^2. We also note that the conditional variable $[\ln S_i|D_i]$ has normal distribution with mean value and variance given by

$$m_{\ln S_i|D_i} = m_S + \rho\frac{\sigma_S}{\sigma_D}[\ln(D_i) - m_D] \quad [S15]$$

$$\sigma^2_{\ln S_i|D_i} = \sigma_S^2(1-\rho^2) \quad [S16]$$

Then $\ln(S_{i,trip}) \sim N(m_S + \rho\sigma_S\sigma_D, \sigma_S^2)$, where $m_S = \ln(S_0 4^{-i}) - \frac{1}{2}\sigma_S^2$. We denote $m_S + \rho\sigma_S\sigma_D$ as μ and σ_S^2 as σ^2. Then μ and σ are the mean and standard error of $\ln(S_{i,trip})$. For a log-normal cascade

$$P_>(i) = \int_1^M (1 - \frac{\zeta}{1-\alpha}(k^{1-\alpha} - 1)) \times \frac{1}{\sqrt{2\pi}\sigma k} e^{-(\ln k - \mu)^2/2\sigma^2} \mathrm{d}k \quad [S17]$$

This integration could be solved:

$$P_>(i) = \frac{e^{-\frac{\mu^2}{2\sigma^2}}(\zeta e^{\frac{(\mu-(\alpha-1)\sigma^2)^2}{2\sigma^2}})\mathrm{erf}(\frac{(\alpha-1)\sigma^2-\mu+\ln x}{\sqrt{2}\sigma})}{2(\alpha-1)}\Big|_{x=1}^{x=M} - \frac{(\alpha-\zeta-1)e^{\frac{\mu^2}{2\sigma^2}}\mathrm{erf}(\frac{\mu-\ln x}{\sqrt{2}\sigma})}{2(\alpha-1)}\Big|_{x=1}^{x=M} \quad [S18]$$

where erf is the error function.

The sensitivity analysis of the influence of different parameters (α in the rank selection mechanism, σ_D, σ_S, and ρ in the cascade model) on the tile exceedance probability $P_>(i)$ is shown in Fig S23. Each time we change one parameter value and fix other parameter values to their calibrated values in Table S2. In Fig S23 (a, b), we change the standard deviation of the log-normal cascade σ_D, σ_S from 0.3 to 0.9. As is shown in Table S2, S and D have positive correlations, so large σ_D will cause trip origins to concentrate in tiles with high supply S, which causes the tile exceedance probability to decrease. The standard deviation of the demand σ_D has a more significant influence on $P_>(i)$. Fig S23 (c, d) shows that $P_>(i)$ is most sensitive to the rank selection parameter $P(k) \sim k^{-\alpha}$. Smaller α means people are less sensitive to trip distance, resulting in longer trip distance and higher tile exceedance probability. Fig S23 (e, f) shows that when the correlation coefficient between the demand and the supply ρ changes from -1 to 1, the level 4 tile exceedance probability changes from 80% to 60%. Negative correlation causes the separation between trip origins and destinations, which increases trip distance.

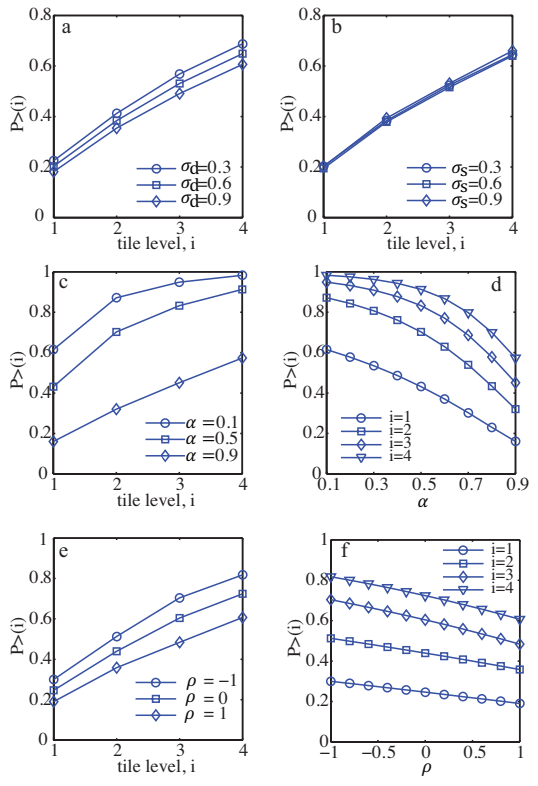

Fig. S23. Sensitivity analysis of different parameters on $P_>(i)$

6 Analytical Derivation of $P(N)$ and $P(\Delta t)$

6.1 Derivation of $P(N)$, the number of daily visited locations.

The power of the Markov framework is that we can analytically derive some of the observed features. Although in the long term observation people could return to any previously visited location, in the calculation in $P(N)$ as a simplification we assume in one day a person could only return to her home and work location (any *other* location will be visited only once in one day). This assumption has been validated in [18].

We subdivide the whole day into T time slots and define a state space of the Markov chain Y_t as

$$\Omega = \{(x,i); x = 0, ..., T-1; i = 0, 1\} \quad [S19]$$

where x denotes how many different locations were already visited and i denotes whether a person is at home ($i = 0$) or away ($i = 1$). For example, $(2, 1)$ means the person has visited 2 locations in this day and now she is not at home. Then it could be shown that the probability of finding our system in the state $Y_T = N$ (i.e. N different locations were visited in T time slots) is:

$$P(Y_T = N) = \xi_0(\prod_{t=1}^{T}\Lambda_t)U^T(C_N) \quad [S20]$$

with ξ_0 being the initial condition (we always start at home) and $U = \sum_{r:a_r \in c_N} U_r$ where U_r is a vector of size $2(T-1)$ with all values zeros except unity at the place corresponding to a state a_r belonging to the subspace $C_N = \{(N, 0), (N, 1)\}$, Λ_t is a time dependent transition probability matrix, which is calculated according to the following rules: $(T-1, 1) \to (T-1, 1)$ and $(T-2, 0) \to (T-2, 0)$ are absorbing states; otherwise the transition probability can be generated from the Markov state transition diagram. For example for $T = 5$ the initial condition vector is $\xi_0 = (1, 0, 0, 0, 0, 0, 0, 0)$ and $U = (0, 0, 0, 0, 0, 1, 1, 0)$ for $Y_5 = 3, (C_N = (3, 0), (3, 1))$. To simplify the notation, we define $n_w P(t)$ as $\mathcal{P}(t)$. The transition probabilities Λ_t are:

$$\begin{aligned}
&(0,0) \to (0,0) : 1 - \mathcal{P}(t)\\
&(0,0) \to (1,1) : \mathcal{P}(t)\\
&(1,0) \to (1,0) : 1 - \mathcal{P}(t)\\
&(1,0) \to (2,1) : \mathcal{P}(t)\\
&(1,1) \to (1,0) : \beta_1\mathcal{P}(1 - \beta_2\mathcal{P}(t))\\
&(1,1) \to (1,1) : 1 - \beta_1\mathcal{P}(t)\\
&(1,1) \to (2,1) : \beta_1\mathcal{P}(t)\beta_2\mathcal{P}(t)\\
&(2,0) \to (2,0) : 1 - \mathcal{P}(t)\\
&(2,0) \to (3,1) : \mathcal{P}(t)\\
&(2,1) \to (2,0) : \beta_1\mathcal{P}(t)(1 - \beta_2\mathcal{P}(t))\\
&(2,1) \to (2,1) : 1 - \beta_1\mathcal{P}(t)\\
&(2,1) \to (3,1) : \beta_1\mathcal{P}(t)\beta_2\mathcal{P}(t)\\
&(3,1) \to (3,0) : \beta_1\mathcal{P}(t)(1 - \beta_2\mathcal{P}(t))\\
&(3,1) \to (3,1) : 1 - \beta_1\mathcal{P}(t)\\
&(3,1) \to (4,1) : \beta_1\mathcal{P}(t)\beta_2\mathcal{P}(t)\\
&(3,0) \to (3,1) : 1; (4,1) \to (4,1) : 1
\end{aligned}$$

We can use $P(Y_T = N) = \xi_0(\prod_{t=1}^{T}\Lambda_t)U^T(C_N)$ to calculate the probability to visit N locations in a day.

6.2 Derivation of $P(\Delta t)$, the stay duration distribution. Similarly, we can also calculate the stay time distribution. In the proposed Markov model, stay duration is the length of consecutively being at a state. For example, in the Markov chain $HHHO_1O_1O_1O_2O_2$, the person first stayed at home for 3 time steps. Then moved to *other* location O_1 and stayed there for 3 time steps. Then moved to *other* location O_2 and stayed there for 2 time slots. In random processes, the length of consecutively being at a state is called *run length*. For the study of consecutive runs of length exactly k (i.e, the stay time is k time slot), we define another finite Markov chain Y_t with state space

$$\Omega = \{(x,i) : x = 0,1,...l; i = -2,-1,0,...,k-1\} - \{(0,-2)\} \quad [S21]$$

A state pair is $Y_t = (x,i)$. T is the total number of time slots; $l = \left\lfloor \frac{T+1}{k+1} \right\rfloor$ is the maximum number of the times that run length k could occur; i means until time step t, in the last $k+1$ time slots, counting from the end of the chain (which is time step t), what is the successive run length. x is from time step 1 to time step t, how many exact run length k has occurred. For example, we want to study run length $k = 3$, then in a chain $HHOOOHHHOHOO$, $x = 2$ since run length 3 occurs twice; $i = 2$ since counting from the last step O, the successive run length is 2. The state $(x,-1)$ and $(x,-2)$ are used for managing the fact that a successive run whose length becomes greater than k does not count. To be more specific:

Overflow state: $(x,-1), x = 0,1,...,l$; which means that in the last $k+1$ time slots, counting from the end of the chain the successive run length is $k+1$. So the run length has already exceeded the target value.

Waiting state: $(x,-2), x = 1,...,l$; which means that in the last $k+1$ time slots, counting from the end of the chain the successive run length is exactly k. So for now this run is regarded as a valid length k run and counted in x. Whether this run could be truly counted depends on the value of the next time step. For example, still for $k = 3$, the chain $OOHHH$ is in state $(1,-2)$. If the next value is O, then we get a run length 3 and the new state is $(1,1)$. On the other hand, if the next value is H, then the run is not valid anymore and the new state is $(0,-1)$. The probability to get j times of run length k in a chain of T time steps is:

$$P(Y_t = j) = \xi_0(\prod_{t=1}^{T} \Lambda_t)U^T(C_j) \quad [S22]$$

ξ_0 is the initial condition. A person always starts at state $(0,0)$. Λ_t is the transition matrix at time step t. $U^T(C_j)$ is the vector of final states we want to get to. For example, if the total number of steps $T = 3$ and we want to find the number of occurrence of run length $k = 2$, the state space is $(0,-1),(0,0),(0,1),(1,-2),(1,-1),(1,0),(1,1)$. There are 7 states in total. The initial condition $\xi_0 = [0,1,0,0,0,0,0]$. If we want to get the probability that run length 2 occurs 0 times, $U^T(C_j) = [1,1,1,0,0,0,0]$. If we want to get the probability that run length 2 occurs 1 time, $U^T(C_j) = [0,0,0,1,1,1,1]$.

Since the probability to travel is different when the person is at "home" or "other" place, these two situations need to be distinguished, which adds one more dimension to the state. Thus $Y_t = \{(x,i,o) : x = 0,1,...l; i = -2,-1,0,...,k-1; o = 0,1\} - \{(0,-2,0),(0,-2,1)\}$. $o = 0$ means the person is at home and $o = 1$ means the person is not at home. Then according to the Markov state transition diagram,the state transition probability at time step t can be written as:

$$P((x,i+1,0),(x,i,0)) = 1 - \mathcal{P}(t), 0 \le x \le l, 0 \le i \le k-2$$
$$P((x,i+1,1),(x,i,1)) = 1 - \beta_1\mathcal{P}(t), 0 \le x \le l, 0 \le i \le k-2$$
$$P((x+1,-2,0),(x,k-1,0)) = 1 - \mathcal{P}(t), 0 \le x \le l-1$$
$$P((x+1,-2,1),(x,k-1,1)) = 1 - \beta_1\mathcal{P}(t), 0 \le x \le l-1$$
$$P((x-1,-1,0),(x,-2,0)) = 1 - \mathcal{P}(t), 1 \le x \le l$$
$$P((x-1,-1,1),(x,-2,1)) = 1 - \beta_1\mathcal{P}(t), 1 \le x \le l$$
$$P((x,-1,0),(x,-1,0)) = 1 - \mathcal{P}(t), 0 \le x \le l$$
$$P((x,-1,1),(x,-1,1)) = 1 - \beta_1\mathcal{P}(t), 0 \le x \le l$$
$$P((x,0,1),(x,i,0)) = \mathcal{P}(t), 0 \le x \le l, -2 \le i \le k-1$$
$$P((x,0,0),(x,i,1)) = \beta_1\mathcal{P}(t)(1-\beta_2\mathcal{P}(t)), 0 \le x \le l, -2 \le i \le k-1$$
$$P((x,0,1),(x,i,1)) = \beta_1\mathcal{P}(t)\beta_2\mathcal{P}(t), 0 \le x \le l, -2 \le i \le k-1$$

The above equations define the transition matrix Λ_t. Then given the initial state and the target final state, the probability to get different stay durations (run length) could be calculated.

1. Candia J et al. (2008) Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical* 41(22):224015.
2. Song C, Koren T, Wang P, Barabási AL (2010) Modelling the scaling properties of human mobility. *Nature Physics* 6(10):818–823.
3. Wang P, Hunter T, Bayen AM, Schechtner K, González MC (2012) Understanding road usage patterns in urban areas. *Scientific reports* 2.
4. Hariharan R, Toyama K (2004) Project lachesis: parsing and modeling location histories in *Geographic Information Science*. (Springer), pp. 106–124.
5. Zheng VW, Zheng Y, Xie X, Yang Q (2010) Collaborative location and activity recommendations with gps history data in *Proceedings of the 19th international conference on World wide web*. (ACM), pp. 1029–1038.
6. Zheng Y, Xie X (2011) Learning travel recommendations from user-generated gps traces. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(1):2.
7. CTPS (2013) Methodology and assumptions of central transportation planning staff regional travel demand modeling.
8. Levinson DM, Kumar A (1994) The rational locator: why travel times have remained stable. *Journal of the american planning association* 60(3):319–332.
9. Schafer A (2000) Regularities in travel demand: an international perspectiv. *Journal of transportation and statistics* 3(3):1–31.
10. U.S. Department of Transportation Federal Highway Administration (2013) CTPP 2006-2010 Census Tract Flows (http://www.fhwa.dot.gov/planning/census_issues/ctpp/data_products/2006-2010_tract_flows/index.cfm).
11. United States Department of Labor Bureau of Labor Statistics (2010) American time use survey (atus), 2010 (http://www.bls.gov/tus/datafiles_2010.htm).
12. Bishop CM (2006) *Pattern Recognition and Machine Learning*. (Springer-Verlag New York, Inc., Secaucus, NJ, USA).
13. Jiang S, Ferreira J, González MC (2012) Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery* 25(3):478–510.
14. U.S. Department of Transportation Federal Highway Administration (2011) 2009 National Household Travel Survey (http://nhts.ornl.gov/download.shtml).
15. Massachusetts Department of Transportation (2012) 2010/2011 massachusetts travel survey. [Online; accessed 17-March-2016].
16. Newman PG, Kenworthy JR (1989) *Cities and automobile dependence: An international sourcebook*.
17. CTPS (2008) Central transportation planning staff regional travel demand modeling methodology and assumptions. [Online; accessed 17-March-2016].
18. 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.