Daily Mobility Patterns - Electronic Supplementary Material

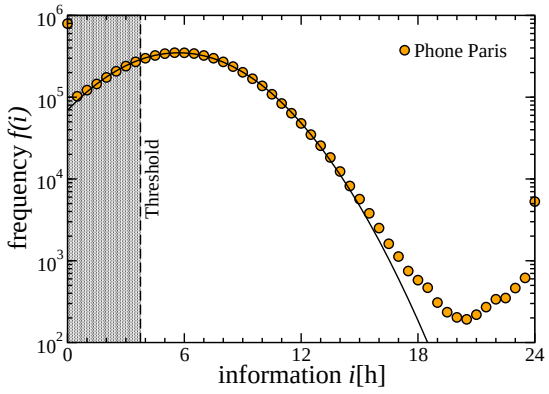
Christian M. Schneider,[1] Vitaly Belik,[1,2] Thomas Couronne,[3] Zbigniew Smoreda,[3] and Marta C. González[1]

[1] *Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA*

[2] *Max Planck Institute for Dynamics and Self-Organization, Am Faßberg, 17, 37077 Göttingen, Germany*

[3] *Sociology and Economics of Networks and Services department, Orange Labs, 38 rue du Général Leclerc, 92794 Issy les Moulineaux, France*

I. MOBILE PHONE DATA PREPROCESSING



Supplementary Figure 1. Number of days with a given number of information.

For our analysis we only take into account days in which a mobile phone user has more information than a certain threshold - 3.5 hours of information distributed in 30-minute intervals, as shown in Supporting Fig. 1. This means that the mobile phone is used at least 8 times during different 30 minute intervals over a single day. The distribution of the number of time intervals with information follows approximately a normal distribution $f(i) = \exp(-(i-\mu)^2/(2\sigma^2))/(\sqrt{2\pi}N\sigma)$ with $\mu = 5.70$ and $\sigma = 3.17$ for $i < 15$ (solid line in Supporting Fig. 1). Interestingly, the distribution has a minimum at $i = 20.5$ and two outliers at $i = 0$, no information, as well as at $i = 24$, full information over the entire day. The threshold is set in such a way that the dataset includes the users with sufficient information. However, a too small threshold would bias the results, since the number of information a user has during a day is the upper limit for the number of visited locations per day. The results for the motif size distribution and the motif distribution itself do not change significantly by changing the threshold to 3 or 4 hours of information.

In Supporting Fig. 2, we illustrate the processing from raw mobile phone data to analyzable data following the rules presented in the Material section. The removed information is marked red. In A) we show the grouping into 48, 30-minute intervals. In B) the towers are removed which are used less than 0.5%. In C) the merging of two towers to one location is shown. In D) non-consecutive locations are identified and removed. In E) an activity is assigned to a location and in F) the home location is added at the beginning (3:30am) and end of the day (3am).

A

User ID	Tower ID	Time	Time Interval
0	0	7:01:32	7am
0	0	7:07:48	7am
0	1	7:15:03	7am
0	2	7:33:59	7:30am
0	3	12:15:15	12pm

B

User ID	Tower ID	Time	Frequency
0	0	7:01:32	65.8%
0	0	7:07:48	65.8%
0	1	7:15:03	0.2%
0	2	7:33:59	13.1%
0	3	12:15:15	0.6%

C

User ID	Tower ID	Time	Location ID
0	0	7:01:32	0
0	1	7:07:48	0
0	0	7:15:03	0
0	1	7:33:59	0
0	0	12:15:15	0
0	1	12:16:10	0
0	0	12:17:34	0

D

User ID	Location ID	Time
0	0	7:01:32
0	0	7:07:48
0	1	7:15:03
0	2	7:33:59
0	3	12:15:15
0	3	12:16:10
0	3	12:17:34
0	2	15:57:15
0	0	16:01:49
0	0	16:05:01
0	2	19:58:34
0	2	20:01:09
0	0	22:46:38
0	0	22:48:20

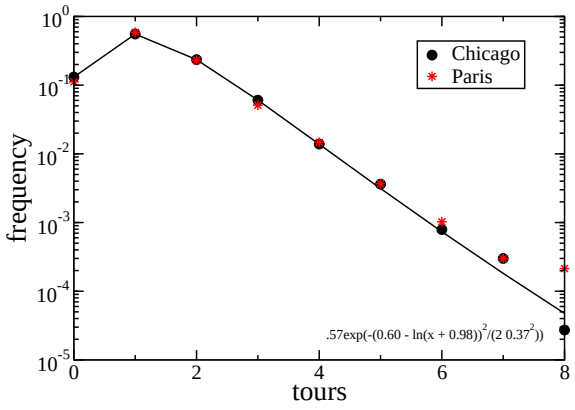
E

User ID	Location ID	Time	Time Interval	Activity ID
0	0	7:01:32	7am	
0	0	7:07:48	7am	0
0	1	7:15:03	7am	
0	2	7:33:59	7:30am	2
0	3	12:15:15	12pm	
0	3	12:16:10	12pm	3
0	3	12:17:34	12pm	

F

User ID	Activity ID	Time Interval
0	0	3:30am
0	0	7am
0	1	7:30am
0	2	12pm
0	2	1pm
0	3	4pm
0	0	5pm
0	0	3am

Supplementary Figure 2. Illustration of the processing of raw data following the rules 1) - 8) presented in the Material section. A) Each call is assigned to the corresponding 30 minute interval. B) Information about infrequently used towers is removed. C) Towers with more than three transitions within a day are merged to one location. D) If less than two consecutive phone activities are performed at a location, this location is ignored. E) In each time interval the activity ID is the most used location ID. F) The first and the last time interval is filled with the most used tower during night time (12pm - 6am).



Supplementary Figure 3. Daily number of tours started at home.

II. TOURS

In Supporting Fig. 3, the number of daily tours is shown for both Chicago and Paris survey starting at home. The distribution can be approximated with $f(N) = C_0 \exp(-(\ln(N+1) - \mu)^2/(2\sigma^2))$ with $C_0 = 0.6$, $\mu = 0.6$ and $\sigma = 0.4$. The broad distribution indicates that although most people perform less than four tours, which are captured with the proposed rules in the paper, some users start up to 8 different tours per day.

III. TRIPS BETWEEN LOCATIONS

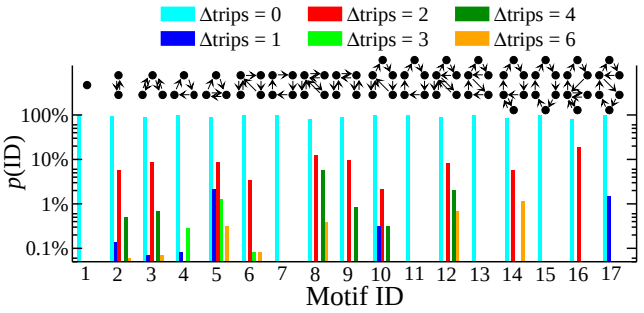
	Chicago Data			Chicago Model		
	Home	Work	Other	Home	Work	Other
Home	<0.01%	12.71%	22.39%	0%	12.45%	22.18%
Work	11.16%	1.94%	5.46%	9.05%	0%	7.25%
Other	24.08%	3.93%	18.33%	25.56%	3.85%	19.63%

Supplementary Table I. Purpose of trips in Chicago.

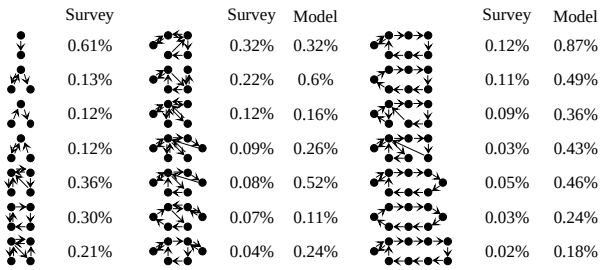
In Supporting Tab. 1, the distribution of trips between home, work and other locations is shown for both Chicago survey and Chicago model. Our proposed model reproduces trips well with an absolute error of at most 2%. The only significant difference is that trips between two different working locations are not present in our model, since it allows only a single work place.

IV. MOTIF INFORMATION

The motifs described in Fig. 3 of the paper are a good proxy for the actual number of performed trips, despite the fact that repetitive trips can not be detected. In



Supplementary Figure 4. Verification that the motifs are a good proxy for the number of actual performed trips. Δtrips is the number of trips which can be removed without changing the corresponding motif of a user.

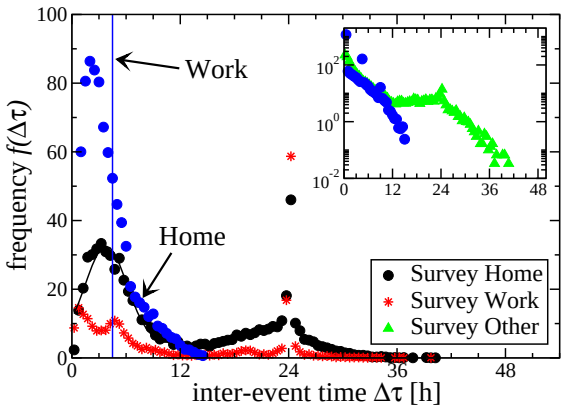


Supplementary Figure 5. Comparison between the most likely insignificant networks in the survey and our model. The model reproduces these networks quite well, if the network fulfill the Eulerian cycle condition.

Supporting Fig. 4, the distribution that a given motif has more than the minimal number of trips is shown for all 17 detected motifs for the Chicago survey data. The minimal number of trips means that a removal of a single trip changes the motif. In $80\% - 100\%$ the minimal number of trips are observed for all motifs. In case a motif contains a tour with only one location $T(1)$, only such tour may be repeated once. The repetition of tours $T(x)$ with $x > 1$ is highly unlikely.

V. INSIGNIFICANT DAILY NETWORKS

The motifs described in Fig. 3 of the paper are not the only daily networks observed in our data. However, if we assume a Poisson distribution for observing networks in the survey, the error of the Chicago survey is $\sqrt{23764} \approx 0.65\%$. All networks below this threshold are statistical not significant. In Supporting Fig. 5 the distribution of the most likely insignificant networks are shown for both survey and the corresponding model. In the survey $3, 37, 74, 102, 83, 35, 21$ additional networks are present for $N = 3, \ldots, 9$ with even lower occurrence probability.



Supplementary Figure 6. Comparison between the intertime distribution of the survey and our model. The blue line indicates the inter-event time for work. The blue circles show the inter-event time between home activities in the main plot and other activities in the inset.

VI. INTER-EVENT TIME COMPARISON

In Supporting Fig. 6 the inter-event time distributions at different places for the survey and our model are compared. Since our model is designed for one day the maximum inter-event time is restricted to $15h$. Due to the fixed working time we observe only a single value at the position of the intraday peak obtained from the survey. The other two inter-event time distributions have the same overall behaviour, a characteristic inter-event time for home activity and fast decaying function for other activities, respectively. However, due to the single short duration spent for other activities the characteristic inter-event time is underestimated and the decay is too fast.

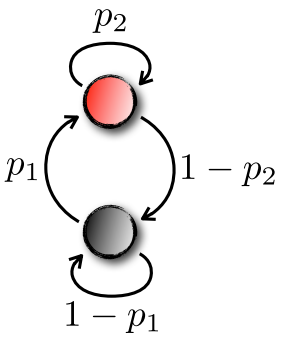
VII. MODELING WORKING AGENTS

Modeling working agents is slightly different from modeling non-working agents. The first difference is that working agents have a second fixed location, their work location, beside the 9h sleeping period. After two free 30-minutes time slots in the morning workers have 16 fixed 30-minutes time slots with a flexible 30-minute time slot in-between for a lunch break. However, during this time, workers prefer to stay at work, if they have no tasks scheduled. Note that tasks which occur during work time are executed during the next free time slot, either during lunch break or after work.
The second difference is a different probability to receive and execute other tasks. Due to the additional work activity, the remaining time for executing other tasks is reduced from 15h for non-working agents to 7h for working agents. Therefore we reduce the probability of a working agent to receive a task p_W, in our simulations p_{NW} is reduced by a factor 1.4:

$$\gamma_W = \gamma_{NW}/1.4 = 1$$

The influence of working agents on the number of visited location N is small, they mainly impact on the probability $p(N)$ for $N = 1$ as shown in Fig. 6b of the paper. However, the existence of two fixed locations explains the occurrence of motif ID 9, which can not be explained with the four proposed rules. Additionally, the results are stable for a wide range of values for $\gamma_W = [0.7 : 1.4]$.

VIII. ANALYTIC SOLUTION

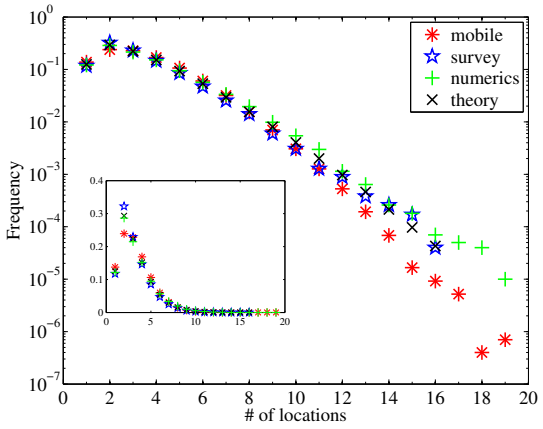


Supplementary Figure 7. Schematic presentation of the human mobility model. Individuals leave home (gray circle) with probability p_1 and remain there with probability $1 - p_1$. If they are not at home (red circle), they move further with a probability $p_2 \gg p_1$ or return home with probability $1 - p_2$.

We model mobility of people as binomial trials – a sequence of successes and failures similar to coin flipping as shown in Supporting Fig. 7. Therefore, we subdivide the whole day into K time slots. A person being home decides in the next time slot to go to another location with probability p_1 (success) and remains home (failure) with probability $1 - p_1$. If the person is already out of home, the probability to visit yet another location is p_2 and the probability to return home is $1 - p_2$.
Next we calculate the distribution of the number of different locations a person visits during one day. For this purpose we use the modified finite Markov chain embedding technique, originally developed for the statistics of success runs in Bernoulli trials.
We define a state space of the Markov chain Y_t as

$$\Omega = \{(x, i);\ x = 0, ..., K - 1;\ i = 0, 1\},$$

where x denotes how many different locations were already visited and i denotes whether a person is at home ($i = 0$) or away ($i = 1$). Then it could be shown that the probability of finding our system in state $Y_K = N$ (i.e.



Supplementary Figure 8. Comparison of the empirical findings for survey and mobile phone data with model predictions – numerical as well as theoretical ones. Inset shows the same plot in linear coordinates.

N different locations were visited during a day) is

$$P\left(Y_K = N\right) = \xi_0 \left(\prod_{t=1}^{K} \Lambda_t \right) U^{\mathsf{T}}(C_N)$$

with ξ_0 being 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 $2(K-1)$ 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 transition probability matrix, which is calculated according to the following rules

$$(K-1,1) \to (K-1,1) \text{ is an absorbing state,}$$
$$(K-2,0) \to (K-2,0) \text{ is an absorbing state,}$$

otherwise

$$(N,0) \to (N,0) \text{ with probability } 1-p_1,$$
$$(N,0) \to (N+1,1) \text{ with probability } p_1,$$
$$(N,1) \to (N,0) \text{ with probability } 1-p_2,$$
$$(N,1) \to (N+1,1) \text{ with probability } p_2.$$

For example for $K = 4$ 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 $Y_4 = 3$ ($C_N = \{(3,0),(3,1)\}$). In this case the transition probability matrix $\Lambda = \Lambda_t$ is given by:

	(0,0)	(1,0)	(1,1)	(2,0)	(2,1)	(3,0)	(3,1)	(4,1)
(0,0)	$1-p_1$		p_1					
(1,0)		$1-p_1$			p_1			
(1,1)		$1-p_2$			p_2			
(2,0)				$1-p_1$			p_1	
(2,1)				$1-p_2$			p_2	
(3,0)						1		
(3,1)						$1-p_2$		p_2
(4,1)								1

To calculate the distribution of visited places, we assign different probabilities p_1 and p_2 to working and non-working agents. In case of working agents an additional location, the work location, is added. Then these two types of agents are mixed with the same ratio as reported in our survey. Finally, we end up with the number of visited locations as shown in Supporting Fig. 8, which can reproduce the obtained results from our surveys well.