Evidence for a Conserved Quantity in Human Mobility

Laura Alessandretti[a], Piotr Sapiezynski[b], Sune Lehmann[b,c,*], and Andrea Baronchelli[a,**]

[a]City, University of London, London EC1V 0HB, United Kingdom

[b]Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

[c]Niels Bohr Institute, University of Copenhagen, DK-2100 København Ø, Denmark

Corresponding authors: *sune.lehmann@gmail.com, **a.baronchelli.work@gmail.com

Abstract

Faced with effectively unlimited choices of how to spend their time, humans are constantly balancing a trade-off between exploitation of familiar places and exploration of new locations. Previous analyses have shown that at the daily and weekly timescales individuals are well characterized by an activity space of repeatedly visited locations. How this activity space evolves in time, however, remains unexplored. Here we analyse high-resolution spatio-temporal traces from 850 individuals participating in a 24-month experiment. We find that, although activity spaces undergo considerable changes, the number of familiar locations an individual visits at any point in time is a conserved quantity. We show that this number is similar for different individuals, revealing a substantial homogeneity of the observed population. We point out that the observed fixed size of the activity space cannot be explained in terms of time constraints, and is therefore a distinctive property of human behavior.

There is a disagreement between the current scientific understanding of human mobility as highly predictable and stable over time [1, 2, 3] and the fact that individual lives are constantly evolving due to changing needs and circumstances [4]. The role of cultural, social and legal constraints on the space-time fixity of daily activities has long been recognized [5, 6, 7]. Recent studies based on the analysis of human digital traces including mobile phone records [8, 9], online location-based social networks [10, 11, 12, 13, 14] and GPS location data of vehicles [15, 16, 17, 18, 19, 20] have shown that individuals universally exhibit a markedly regular pattern characterized by few locations where they return regularly [21, 22] and predictably [23]. However, the observed regularity mainly concerns human activities taking place at the daily [24, 25] or weekly [11, 8, 9] time-scales, such as commuting between home and office [8, 9, 26, 27], pursuing habitual leisure activities, and socializing with established friends and acquaintances [10]. The role played by slowly occurring life changes is not well understood and their effects are not included in the available models of human mobility behavior [2, 28, 29, 30, 31, 32].

Here, we quantify the development of individuals routines across months and years, characterising how individuals balance the trade off between the exploitation of familiar places and the exploration of new opportunities. Our study is based on the mobility traces of ~ 850 university students involved in the Copenhagen Networks Study experiment [33, 34] over a period of 24 months (Fig. S1A). The physical

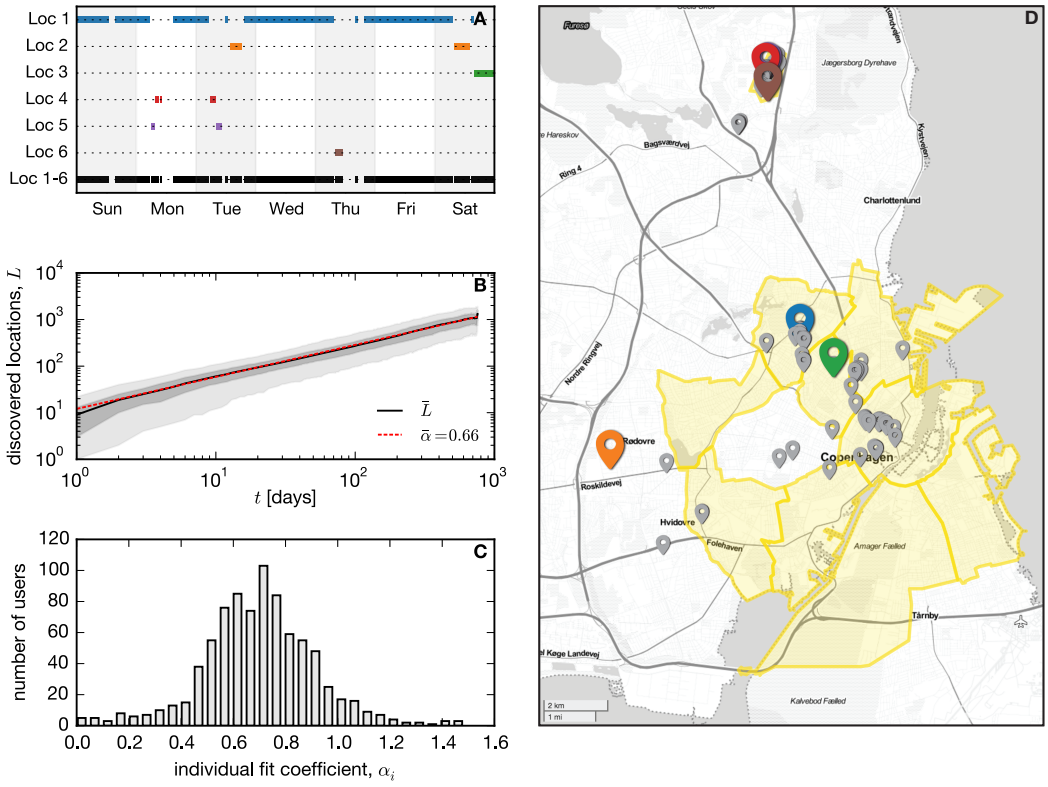

Figure 1: Activity space and exploration of new locations. (**A**) An example of individual mobility trace. The visiting temporal pattern of the six most visited locations are shown (Loc1, ..., Loc6) along with the black trace including all visits to these 6 locations (Loc1-6). (**B**) Total number of discovered locations, L. Time is measured in t days since an individual received the phone. The figure shows the 50% (dark grey area) and the 90% (light gray area) of the population, the average across users $\overline{L}$ (black line) and a power-law fitting function (red dashed line) with exponent $\overline{\alpha}$. (**C**) The distribution of individuals power-law fit coefficients α_i is peaked around its average value $\overline{\alpha} = 0.66$. (**D**) Example of an individual's activity space. Locations are represented as pins on a map. The six most visited locations are displayed as larger pins using the same color scheme of panel A. Yellow areas show Copenhagen city and DTU University.

location of individual mobile devices across time is inferred using a combination of an individual's WiFi scan time-series and GPS coordinate scans (see SI, Figs. S1C and S2). The temporal sampling of WiFi traces in the dataset is even, with median time between scans $\Delta t = 16sec$ (Fig. S1B), while ubiquitous WiFi access points (AP) can be localized with a typical spatial resolution of the order of tens of meters [35, 36]. Fixed rate sampling and high spatial resolution allow us to capture mobility patterns beyond highly regular ones such as home-work commuting [37], avoiding possible biases in the analyses based on location data extracted from mobile phone calls and location-based social networks typically used for mobility analyses [26].

Importantly, notwithstanding significant differences in data collection and the homogeneity of the sampled population (university students), we have verified that the Copenhagen Networks Study dataset displays statistical properties that are consistent with previously analysed data on human mobility across a wide range of measures [1, 2] (Fig. S3). Yet its temporal duration and spatial resolution makes it ideal for investigating the evolution of individual geo-spatial behaviors on longer timescales.

Results

When initiating a transition from a place to another, individuals may either choose to return to a previously visited place, or explore a new location. To characterize this exploration-exploitation trade-off, we represent individual geo-spatial trajectories as sequences of locations, where 'locations' are defined as places where participants in the study stopped for more than 10 minutes (SI text, Fig. 1A). A first question concerning the *exploration behavior* of the individuals is whether an individual's set of known locations is continuously expanding, or instead its size saturates over time. We find that the total number of unique locations $L_i(t)$ an individual i has discovered up to time t grows as $L_i \propto t^{\alpha_i}$ (Fig. 1B), and that individuals' exploration is homogeneous across the population studied, with α_i peaked around $\overline{\alpha} = 0.66$ (Fig. 1C). This sub-linear growth occurs regardless of how locations are defined or when in time the measurement starts (Fig. S4). It is a characteristic signature of Heaps' law [38] and implies that the rate of discovery of new locations decreases continuously during the entire duration of the experiment.

We also find that, while continually exploring new places, individuals allocate most of their time among a small subset of all visited locations (Fig. S5A), in agreement with previous research on human mobility behavior [21, 22, 23] and time-geography [5, 39, 40, 41, 42]. Hence, at any point in time, each individual is characterized by an *activity space* (AS) (Fig. 1D), defined as the subset of all locations within which she visits as a result of her daily activities [40]. Operationally, we define the activity space as the set $AS_i(t) = \{\ell_1, \ell_2, ..., \ell_k, ...\ell_C\}$ of locations ℓ_k that individual i visited at least twice and where she spent on average more than 10 minutes/week during a time-window of 10 consecutive weeks preceding time t. The results presented below are robust with respect to variations of this definition, such as changes of the time-window size or the definition of a location (Tables S1 - S5, Figs. S4A, S5B, S6).

Thus, individuals continually explore new places yet they are loyal to a limited number of familiar locations forming their AS. But how does discovery of new places affect an individual's AS? We find that the average probability $\overline{P}$ that a newly discovered location will become part of the AS stabilizes at $\overline{P} = 20\%$ over the long term, indicating that individual AS are inherently unstable and new locations are continually added. However, over time individuals may also cease to visit locations that are part of the AS. The balance between newly added and dismissed familiar locations is captured by the temporal evolution of the AS, which we characterize by the *spatial capacity* and *net gain*. We define *spatial capacity* C_i as the number of an individual's familiar locations, i.e. the AS size, at any given moment. The *net gain* G_i is defined as the difference between the number of locations that are respectively added (A_i) and removed (D_i) at a specific time, hence $G_i = A_i - D_i$. Fig. 2A shows the evolution of the average population capacity $\overline{C}$. We find that the average capacity $\overline{C}$ is constant in time, with a linear fit of the form $\overline{C} = a + m \cdot t$ yielding $m = 0.05 \pm 0.1$. Analogously, a power-law fit of the form $\overline{C(t)} \propto t^{\beta}$ yields $\beta = -0.03 \pm 0.07$. As a further control, we performed a multiple hypothesis test with false discovery rate correction to compare the averages of the capacity distribution at different times (Table S6). We find no evidence for rejecting the hypothesis that the average capacity does not change in time. Thus, despite individual AS evolving over time, the average capacity is a conserved quantity.

The conservation of the average spatial capacity may result from either (i) each individual maintaining a stable number of familiar locations over time or (ii) a substantial heterogeneity of the population with certain individuals shrinking their set of familiar locations and other expanding theirs. We test the two hypotheses by measuring the individual average net gain across time $\langle G_i \rangle$ and its standard deviation $\sigma_{G,i}$. If a participant's average gain is closer than one standard deviation from 0, hence $|\langle G_i \rangle|/\sigma_{G,i} < 1$, then

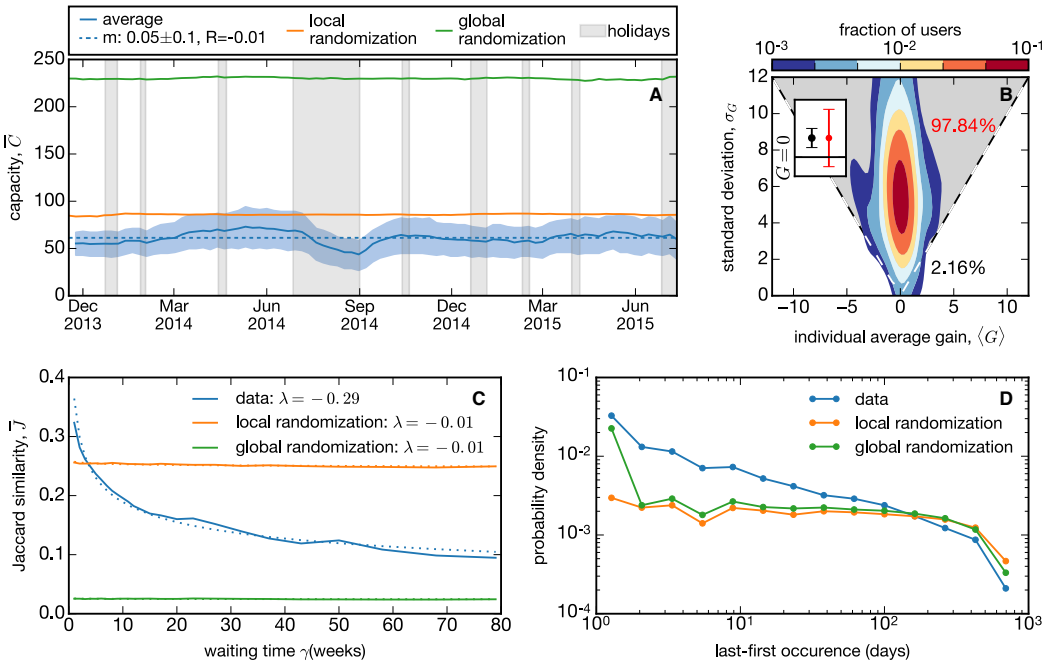

Figure 2: Conserved size of evolving activity spaces. (**A**) Evolution of individual capacity. The light blue area represents 50% of the population, the blue line its average $\overline{C}$, and the blue dashed line is the linear fit. The error on the angular coefficient m of the fitting line, reported in the legend, shows that the fit is compatible with a constant line. The capacities resulting from the local (orange line) and global (green line) randomizations are also reported. (**B**) Gain standard deviation $\sigma_{G,i}$ vs the average gain $\langle G_i \rangle$ (lines obtained through a kernel density estimation from the data). The grey area corresponds to individuals for which $|\langle G_i \rangle| < \sigma_{G,i}$, i.e. whose average gain is compatible with zero (red bar in the illustrative inset). It contains 97.84% of the population. (**C**) The average Jaccard similarity $\overline{J}$ between the weekly activity spaces measured at t and $t+\gamma$ as a function of γ for data (blue line), and the randomized series (orange and green lines). Dashed lines correspond to power-law fits $\overline{J} \sim \gamma^{\lambda}$. (**D**) Probability distribution of the time interval between first and last occurrences of a location for data (blue line) and the randomized cases (orange and green lines).

the net gain is consistent with $\langle G_i \rangle = 0$. If this is true for the majority of participants, the spatial capacity is conserved at the individual level and hypothesis (i) holds. If, on the other hand, $|\langle G_i \rangle|/\sigma_{G,i} \geq 1$, the individual capacity must either increase or decrease in time, supporting hypothesis (ii). We find that hypothesis (i) holds for for 97.84% of individuals (Fig. 2B, see SI text for further analysis of this point). For the large majority of the population, the average net gain of familiar locations added or removed to the AS at any instant of time is not significantly different from 0, hence their individual capacity is conserved (Table S4). Also, we find that the ratio between the average individual capacity and its standard deviation across time is smaller than 23% for 75% of the population (Fig. S7C), demonstrating that fluctuations of the capacity are relatively small.

These results indicate that each individual is characterized by a fixed-size but evolving set of familiar locations. While the size depends on the spatial resolution chosen to define locations (Fig.S7, caption), the population is homogeneous within this constraint (Fig. 3B). To interpret the information contained in the precise value of the spatial capacity, we randomize the temporal sequences of locations in two ways, preserving routines of individuals only up to the daily level. After breaking individual time series into modules of 1 day length, (a) we randomize individual timeseries preserving the module/day units (local randomizations) or (b) we create new sequences by assembling together modules extracted randomly by

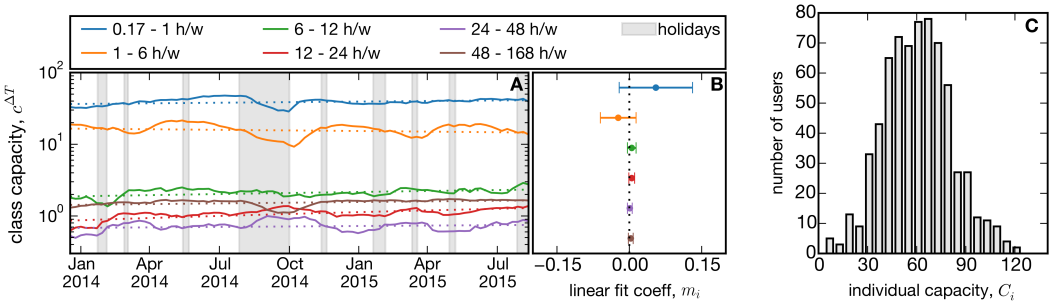

Figure 3: Conservation of time allocation. (**A**) The average capacities (full lines) $c^{\Delta T}$ and the corresponding fits (dashed lines) as a function of time computed for several categories of locations ΔT. (**B**) Fit coefficients are consistent with 0 within errors (Table S8). (**C**) Frequency histogram of individuals according to their average individual capacity C_i.

the whole set of individual traces (global randomization) (Fig S8A-C). Due to the absence of temporal correlations, the capacity is constant in time also for the randomized datasets (see Fig. 2A). However, the capacity of the random sets is significantly higher than in the real time series for both randomizations (Fig. S8D-F, Table S7), implying that the observed value in real data is not a simple consequence of time constraints. Instead, the fixed capacity is an inherent property of human behavior.

The time evolution of the AS supports this finding. We measure the turnover of familiar locations using the Jaccard similarity $J_i(t,\gamma)$ between the weekly AS at t and at $t+\gamma$ (Fig. 2C). Despite seasonality effects which imply fluctuations around a typical behavior, J_i does not depend on the initial point but only on the waiting time γ, and we can consider $J_i(\gamma)$ independently of t (Fig. S5C). We find that the average similarity decreases as a power law $\overline{J} \propto \gamma^{\lambda}$ with coefficient $\lambda = -0.29$. On the other hand, for the randomized sequences, the Jaccard similarity is constant in time as familiar locations are never abandoned ($\overline{J} \propto \gamma^0$). Also, individuals keep visiting only few locations for long periods of time, in contrast to the randomized cases (Fig. 2D). This confirms that individual sets of familiar locations change continually and individual routines evolve gradually in time.

In order to characterize the structure of the activity space, we investigate how individuals allocate time among different classes of locations defined on the basis of their average visit duration. We consider intervals ΔT, with ΔT ranging from 10 to 30 minutes per week (the time it takes to visit a bus stop or grocery shop) up to 48 to 168 hours per week (such as for home locations) (Fig. S9).

For each of these locations classes, we compute the evolution of the *capacity* $c_i^{\Delta T}$ and the *gain* $G_i^{\Delta T}$, and test the hypothesis $G_i^{\Delta T} = 0$, as above. We find that, although the AS subsets are continuously evolving (Fig. S9G), $c_i^{\Delta T}$ is conserved for each ΔT (Fig. 3C, Tables S5 and S8), indicating that the number of places where individuals spend a range of time ΔT does not change over time. This result holds independently of the choice of specific ΔT (Table S8) and implies that the individual capacity $C_i = \sum c_i^{\Delta T}$, where both C_i and each $c_i^{\Delta T}$ are conserved across time. Thus, both spatial capacity and time allocation are conserved quantities.

Discussion

In summary, we have shown that the number of locations an individual visits regularly is conserved over time, even while individual routines are unstable in the long term because of the continual exploration of new locations. This individual *spatial capacity* is peaked around a typical value across the population,

which is significantly lower than expected if only time-constraints were at play. Finally, this spatial capacity is hierarchically structured, indicating that individual time allocation for categories of places is also conserved.

Taken together, these findings shed new light on the underlying dynamics shaping human mobility, with potential impact for a better understanding of phenomena such as urban development and epidemic spreading. They will also help test and improve existing models of human mobility[2, 28, 29, 30, 31, 32] which were not designed to account for long-term instabilities and fixed-capacity effects. Extending our scope beyond mobility, it is interesting to note that similar fixed-size effects in the social domain [43, 44, 45, 46] have been put in direct relation with human cognitive abilities [43]. We anticipate that our results will stimulate new research exploring this connection.

Data description

The data was collected by the Copenhagen Network Study (CNS) experiment [33]. The CNS data collection took place between September 2013 and September 2015. In total, 851 students were involved in the experiment. Data collection was approved by the Danish Data Protection Agency. All participants provided informed consent.

We estimate participants position over time by combining WiFi scans data and GPS scans data. The WiFi dataset provides the time-series of wireless network scans performed by participants' mobile devices. Each record *(i, timestamp, SSID,BSSID, RSSI)* indicates the participant ID i, the name of the wireless network scanned *SSID*, the MAC address *BSSID* uniquely identifying the AP providing access to the wireless network, the time of scan in seconds *timestamp*, and the signal strength *RSSI* in dBm. APs do not have geographical coordinates attached. However, their time sequence is effectively equal to location data, because APs positions tend to be fixed.

References

[1] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[2] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

[3] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[4] Irwin G Sarason, James H Johnson, and Judith M Siegel. Assessing the impact of life changes: development of the life experiences survey. *Journal of consulting and clinical psychology*, 46(5):932, 1978.

[5] Torsten Hägerstraand. What about people in regional science? *Papers in regional science*, 24(1):7–24, 1970.

[6] Lawrence D Burns. Transportation, temporal, and spatial components of accessibility. 1980.

[7] Tim Schwanen, Mei-Po Kwan, and Fang Ren. How fixed is fixed? gendered rigidity of space–time constraints and geographies of everyday activities. *Geoforum*, 39(6):2109–2121, 2008.

[8] Balázs Cs Csáji, Arnaud Browet, Vincent A Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459–1473, 2013.

[9] Andres Sevtsuk and Carlo Ratti. Does urban mobility have a daily routine? learning from the aggregate data of mobile networks. *Journal of Urban Technology*, 17(1):41–60, 2010.

[10] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[11] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z Sui. Exploring millions of footprints in location sharing services. *ICWSM*, 2011:81–88, 2011.

[12] Chloë Brown, Neal Lathia, Cecilia Mascolo, Anastasios Noulas, and Vincent Blondel. Group colocation behavior in technological social networks. *PloS one*, 9(8):e105816, 2014.

[13] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027, 2012.

[14] Halgurt Bapierre, Chakajkla Jesdabodi, and Georg Groh. Mobile homophily and social location prediction. *arXiv preprint arXiv:1506.07763*, 2015.

[15] Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvatore Rinzivillo, and Roberto Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB JournalThe International Journal on Very Large Data Bases*, 20(5):695–719, 2011.

[16] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T Campbell. Nextplace: a spatio-temporal prediction framework for pervasive systems. In *Pervasive Computing*, pages 152–169. Springer, 2011.

[17] Xiao Liang, Xudong Zheng, Weifeng Lv, Tongyu Zhu, and Ke Xu. The scaling of human mobility by taxis is exponential. *Physica A: Statistical Mechanics and its Applications*, 391(5):2135–2144, 2012.

[18] Gallotti Riccardo, Bazzani Armando, and Rambaldi Sandro. Towards a statistical physics of human mobility. *International Journal of Modern Physics C*, 23(09), 2012.

[19] Armando Bazzani, Bruno Giorgini, Sandro Rambaldi, Riccardo Gallotti, and Luca Giovannini. Statistical laws in urban mobility from microscopic gps data in the area of florence. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(05):P05001, 2010.

[20] Bin Jiang, Junjun Yin, and Sijian Zhao. Characterizing the human mobility pattern in a large street network. *Physical Review E*, 80(2):021136, 2009.

[21] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in peoples lives from cellular network data. In *Pervasive computing*, pages 133–151. Springer, 2011.

[22] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature communications*, 6, 2015.

[23] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[24] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[25] James P Bagrow and Yu-Ru Lin. Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5):e37676, 2012.

[26] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3):33–44, 2012.

[27] Hui Zang and Jean Bolot. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, pages 145–156. ACM, 2011.

[28] Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 239–252. ACM, 2012.

[29] Kyunghan Lee, Seongik Hong, Seong Joon Kim, Injong Rhee, and Song Chong. Slaw: A new mobility model for human walks. In *INFOCOM 2009, IEEE*, pages 855–863. IEEE, 2009.

[30] Minkyong Kim, David Kotz, and Songkuk Kim. Extracting a mobility model from real user traces. In *INFOCOM*, volume 6, pages 1–13, 2006.

[31] Tao Jia, Bin Jiang, Kenneth Carling, Magnus Bolin, and Yifang Ban. An empirical study on human mobility and its agent-based modeling. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(11):P11024, 2012.

[32] Xiao-Pu Han, Qiang Hao, Bing-Hong Wang, and Tao Zhou. Origin of the scaling law in human mobility: Hierarchy of traffic systems. *Physical Review E*, 83(3):036117, 2011.

[33] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. Measuring large-scale social networks with high resolution. *PloS one*, 9(4):e95978, 2014.

[34] Vedran Sekara, Arkadiusz Stopczynski, and Sune Lehmann. Fundamental structures of dynamic social networks. *Proceedings of the National Academy of Sciences*, 113(36):9977–9982, 2016.

[35] Piotr Sapiezynski, Radu Gatej, Alan Mislove, and Sune Lehmann. Opportunities and challenges in crowdsourced wardriving. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pages 267–273. ACM, 2015.

[36] M Mun, Deborah Estrin, Jeff Burke, and Mark Hansen. Parsimonious mobility classification using gsm and wifi traces. In *Proceedings of the Fifth Workshop on Embedded Networked Sensors (HotEmNets)*, 2008.

[37] Serdar Çolak, Lauren P Alexander, Bernardo Guatimosim Alvim, Shomik R Mehndiretta, and Marta C González. Analyzing cell phone location data for urban travel: Current 2 methods, limitations and opportunities 3. In *Transportation Research Board 94th Annual Meeting*, number 15-5279, 2015.

[38] Harold Stanley Heaps. *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc., 1978.

[39] Frank E Horton and David R Reynolds. Effects of urban spatial structure on individual behavior. *Economic Geography*, 47(1):36–48, 1971.

[40] Mary Ellen Mazey. The effect of a physio-political barrier upon urban activity space. 1981.

[41] Reginald G Golledge. *Spatial behavior: A geographic perspective*. Guilford Press, 1997.

[42] Yihong Yuan and Martin Raubal. Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study. *International Journal of Geographical Information Science*, 30(8):1594–1621, 2016.

[43] Robin IM Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and brain sciences*, 16(04):681–694, 1993.

[44] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3, 2013.

[45] Jari Saramäki, EA Leicht, Eduardo López, Sam GB Roberts, Felix Reed-Tsochas, and Robin IM Dunbar. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, 111(3):942–947, 2014.

[46] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users activity on twitter networks: Validation of dunbars number. *PloS one*, 6(8):e22656, 2011.

[47] Piotr Sapiezynski, Radu Gatej, Alan Mislove, and Sune Lehmann. Opportunities and challenges in crowdsourced wardriving. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pages 267–273. ACM, 2015.

[48] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[49] Leendert Cornelis Elisa Struik. *Physical aging in amorphous polymers and other materials*. PhD thesis, TU Delft, Delft University of Technology, 1977.

[50] Animesh Mukherjee, Francesca Tria, Andrea Baronchelli, Andrea Puglisi, and Vittorio Loreto. Aging in language dynamics. *PLoS One*, 6(2):e16677, 2011.

Evidence for a Conserved Quantity in Human Mobility - Supplementary Information

1 Data pre-processing

Access Points (AP) locations were estimated using participants' sequences of GPS scans. First, we discarded *mobile APs*, that are located on buses or trains, and *moved APs* that were displaced during the experiment (for example by residents of Copenhagen changing apartment, taking their APs with them). Then, we considered all WiFi scans happening within the same second as a GPS scan to estimate APs location. The APs location estimation error is below 50 meters in 99% cases. Most of the APs are located in the Copenhagen area (Fig. 1C and [47] for a detailed description of the methodology).

Quantifying individuals' mobility behavior requires a definition of spatial "locations" identifying places that are significant in human daily experience (i.e. homes, offices, cafés, shops...). There is not an unequivocal definition of "locations": here, we cluster APs into "locations" based on the distance between them. The information on simultaneous detection of two APs is included in the indirect graph $G = (V, E)$. V is the set of geo-localised APs, links $e(j, k)$ exist between pairs of access points that have ever been scanned in the same 1 min bin by at least one user. The physical distances $dist(j, k)$ for all pairs of $(j, k) \in E$ can be easily computed since their position has been estimated. We consider the set of links $E_D \subset E$ such that $dist(j, k) < d$, where d is a threshold value, and we define a new graph $G_d = (V, E_d)$. Each connected component in the graph G_d include all APs that are closer than d to at least one other AP in the same component. With our definition, a *stop-location* is a connected component in the graph G_d. For $d = 5m$ the maximal distance between two APs in the same location is smaller than $10m$ for most locations and at most $\sim 200m$ (Fig. S2A). The number of APs in the same location is lower that 10 for most locations, but reaches ~ 700 for dense areas such as the University Campuses (one should also consider that APs are replaced throughout the experiment) (see Fig. S2B). An example of APs clustering for $d = 5m$ and $d = 10m$ is shown in Fig. S2C and D. Our findings do not depend on the choice of the threshold.

Throughout the experiment, participants' devices scanned for WiFi every Δt seconds. For half of the population the median time between scans is lower than $\Delta t_m = 16$ sec (Fig. S1B). This high resolution temporal granularity is not meaningful for the study of individual long-term behavior. Data was aggregated in bins of length 1min. Individuals long-term mobility behavior can be estimated for individuals whose position is known for a considerable fraction of time. For each individual i, we measure the *time coverage* TC_i as the fraction of time an individual's location is known. The time coverage displays fluctuations due to seasonality effects (Fig. S10A), but is higher than 0.8 for 50% of the population (Figs. S10B and C) when considering only geo-localised APs.

2 Comparison with previous research

Our dataset displays statistical properties that are consistent with previously analysed data on human mobility. The distribution between consecutive jumps for $\Delta r \geq 1$ Km can be modelled as $P(\Delta r) \sim \Delta r^{\beta}$ (Fig. S3A), with exponent $\beta = -1.87$. Gonzales et. al [1] found $\beta = -1.75 \pm 0.15$ for the truncated

power-law distribution, Song et. al [2] found a power-law distribution of jumps with exponent $\beta = -1.55$. Individuals are distributed heterogeneously with respect to their radius of gyration (see [1], SI for definition) measured at the end of the experiment, with the probability distribution $P(r_g)$ (Fig. S3B) decaying as a power-law $P(r_g) \sim r_g^{\beta}$ with coefficient $\beta = -1.47$. This is comparable with the results found in [1], $\beta = -1.65$ and [2] $\beta = -1.55$, where both studies relied on CDRs.
The visitation frequency of a location is defined as $f_l = m_l / \sum_{k=0}^{NL}(m_k)$ where m_l is the total number of visits to the $l-th$ location, and NL is the number of locations. We find that the visitation frequency of a location with rank r, where the rank is attributed based on the visitation frequency, follows a Zipf's law $f(r) \propto r^{\zeta}$, with $\zeta = -1.2$ (Fig. S3C). Our result is consistent with the one obtained by Song et al who found $f(r) \propto r^{-1.2 \pm 0.1}$[2]. Gonzalez et. al found that $f(r) \propto 1/r$ [1].

3 Robustness Tests

The results presented in the main text do not depend on how locations are defined (most results hold simply considering Wi-Fi access points (APs) as proxy for locations), nor on the time-window used to investigate the long-term behavior. In this section, we show how the results are derived and we demonstrate their statistical robustness. To avoid confusion, we will indicate with $\overline{x}$ the average value of a quantity x across the population, and $\langle x \rangle$ the average across time.

Exploration grows sub-linearly regardless of the definition of location

Individual exploration behavior is quantified measuring the number of locations $L_i(t)$ discovered up to day t. In the MS, we show that for $d = 5m$, individual exploration does not saturate, with $L_i(t)$ growing sub-linearly in time. Here, we show that this holds also considering APs and locations with threshold $d = 10m$ (Fig. S4A).

Conservation of capacity

The activity space is defined here as the set $AS_i(t) = \{\ell_1, \ell_2, ..., \ell_k, ...\ell_C\}$ of locations ℓ_k that individual i visited at least twice and where she spent on average more than 10 minutes/week during a time-window of W consecutive weeks preceding time t. Given this definition, the number of locations an individual i visits regularly is equivalent to the activity space size $C_i(t) = |AS_i(t)|$. We call this quantity *spatial capacity*. The average individual capacity across the population $\overline{C(t)}$ is constant in time regardless of the choice of the window size W (Table S1 and Table S2) or the definition of location. This is tested by first performing a linear fit of the form $\overline{C(t)} = a + m \cdot t$ and a power-law fit of the form $\overline{C(t)} \propto t^{\beta}$, and then testing the hypotheses $H_0 : m = 0$, and $H_1 : \beta = 0$. In Table S1 we show the fit coefficients computed with the least squares method, where the errors are estimated taking into account the standard deviation $\sigma_C(t)$ quantifying the population dispersion. Hypotheses H_0 and H_1 hold for all the choices of W. Also, the correlation hypothesis test, testing $\rho_{t,C} = 0$, where $\rho_{t,C}$ is the Pearson correlation coefficient between capacity and time, yields that there is not significant correlation between time and capacity at $\alpha = 0.05$ with p-value$> \alpha$ for any choice of W (Table S1). The coefficient of determination R^2 is smaller than 0 in most cases meaning that a horizontal line fits better than the best fit computed with the least squared method.

As an additional test, we verify that the capacity is constant by comparing its average value across different time-intervals. We divide the total time range into time-intervals δt_k spanning W weeks. We compute the average capacity $\overline{C(\delta t_k)}$ and its standard deviation $\sigma_C(\delta t_k)$ for each time-interval δt_k as:

$$\overline{C(\delta t_k)} = \frac{1}{W} \sum_{t' \in \delta t_k} \overline{C(t')}$$

$$\sigma_C(\delta t_k) = \sqrt{\frac{1}{W} \sum_{t' \in \delta t_k} \sigma_C(t')^2}$$

We test the hypothesis $H_{j,k} : \overline{C(\delta t_k)} = \overline{C(\delta t_j)}$ for all pairs $\delta t_k, \delta t_j$. The hypotheses are tested using independent 2-samples t-tests, testing if two populations with the same variance have equal mean, (using the mean computed from randomly drawn samples). The test statistics $t_{k,j}$ for the student's t distribution with $df_{k,j} = 2 * w - 2$ degrees of freedom is computed as :

$$t_{k,j} = \frac{\overline{C(\delta t_k)} - \overline{C(\delta t_j)}}{\sqrt{\frac{1}{W}(\sigma_C(\delta t_k)^2 + \sigma_C(\delta t_j)^2)}}$$

We compute the p-value $p_{j,k}$ (the probability of observing a greater difference between $\overline{C(\delta t_k)}$ and $\overline{C(\delta t_j)}$ given that the two populations have the same mean) for all intervals (Table S6). We apply a False Discovery Rate controlling procedure [48] for multiple testing to control the proportion of rejected null hypotheses that were incorrect rejections. Applying this methodology we find that for all j, k, the hypothesis $H_{j,k} : \overline{C(\delta t_k)} = \overline{C(\delta t_j)}$ can not be rejected at $\alpha = 0.05$ (results hold both for locations and APs, for locations with threshold d=5 see Table S6). This result holds considering all windows sizes included between $W = 2$ and $W = 10$ weeks.

The individual average capacities $\langle C_i \rangle$ are homogeneously distributed around a typical value for the population that is determined by the choice made to define location (Fig. S7A). The capacity of a single individual, $C_i(t)$, is also conserved across time. We apply a linear fit of the form $C_i(t) = a_i + m_i \cdot t$ and find that the sample mean of the linear fit coefficients is $\overline{m} = -0.054 \pm 0.048$ (Fig. S7B and Table S3). The hypothesis $\overline{m} = 0$ is not rejected at $\alpha = 0.05$ under the t-student hypothesis test, for locations with $d = 5m$, and APs (Table S3). We quantify the fluctuations of the individual capacity measuring the coefficient of variation $\sigma_{C,i}/\langle C_i \rangle$, where $\langle C_i \rangle$ is an individual's average capacity and $\sigma_{C,i}$ the corresponding standard deviation. For $W = 10$, these fluctuations are smaller than 23% for 50% of individuals (Fig. S7C), considering only weeks with time coverage higher than 80%.

Changes of the individual capacity are also quantified measuring the *net gain*, defined as $G_i(t) = A_i(t) - D_i(t)$, where $A_i(t) = |AS_i(t) \setminus AS_i(t - dt)|$ is the number of location added and $D_i(t) = |AS_i(t - dt) \setminus AS_i(t)|$ (the difference between the sets) is the number of location removed from the activity space during dt, where $dt = 1$ week. Fig. S6 shows the individual average capacity $\langle C_i \rangle$, activation $\langle A_i \rangle$, deactivation $\langle D_i \rangle$ and weekly gain $\langle G_i \rangle$, as a function of the windows size W. The individual average gain $\langle G_i \rangle$ is consistent with 0 for all choices of W, because of the conservation of individual capacities across time (Fig. S6, right column). The individual average gain $\langle G_i \rangle$ is consistent with 0 for more than 95% of individuals, independently of how locations are defined. This is verified in 2 ways:

- We test whether the ratio $\sigma_{G,i}/\langle G_i \rangle > 1$, where $\sigma_{G,i}$ is the standard deviation of the average

individual net gain across time (see main text). In Table S4, we show the correlation between $\sigma_{G,i}$ and $\langle G_i \rangle$, for different definitions of locations.

- We compute the 95% confidence interval $\langle G_i \rangle \pm 1.96 * \sigma_{G,i}/\sqrt{n_t}$, where n_t is the number of data points for each individual (i.e. the number of weeks she joined the experiment). We find that the confidence interval includes 0 for over 95% of individuals in the experiment, for all definitions of locations and window sizes.

Finally, we verify that the capacity is conserved using a different methodology for inferring individual trajectories from WiFi data. Here, we discover the routers' locations using the approach described in [47] with a slight modification. The original method used only GPS location estimations calculated at the same second as a corresponding WiFi scan. Here, we consider all location estimations from Android Location API, including network based estimations. Additionally, we relax the same-second requirement as follows. In the spatio-temporal trace of each user we identify periods from time t_0 to time t_N where the user was stationary, also referred to as stop locations. This means that the distance between the user's location at t_0 and t_N is below d meters, and that there exist a location estimation between t_0 and t_N at least every n seconds. Also, it implies that each location estimation within the stop location is within d from the user's location at t_n. We select n as 305 seconds, thus requiring no missing data — the sampling period of location in the experiment is approximately 300 seconds. We select d as 30 meters, thus requiring that the user remains in the same location within the resolution of a building. After identifying these stop locations, we assign the median position of estimations to all routers scanned in these periods. Then, we follow the procedure described in [47], and we cluster routers based on distance as previously described. We verify that capacity is conserved under the same tests described above also in this case. The linear test fit gives $m = 0.02 \pm 0.07$.

Invariance of exploration behavior under time translation

Here we verify that all the measures we calculated are not influenced by the particular time at which the data collection started or by the time elapsed from that moment. We borrow the concept of *aging* from the physics of glassy systems [49, 50]. A system is said to be in equilibrium when it shows invariance under time translations; if this holds, any observable comparing the system at time t with the system at time $t + \gamma$ is independent of the starting time t. In contrast, a system undergoing aging is not invariant under time translation. This property can be revealed by measuring correlations of the system at different times.

We verify that exploration behavior is not affected by the starting time as we verify by measuring the locations discovered starting the measure M months after the participant received the phone, with M in {1,2,5,7,10,12} (Fig. S4B).

Also, we measure the evolution of the activity space starting at different initial times t to verify if the system undergoes aging effects. The evolution is quantified measuring the Jaccard similarity $J_i(t, \gamma) = |AS_i(t) \cap AS_i(t + \gamma)|/|AS_i(t) \cup AS_i(t + \gamma)|$ (Fig. S5B). The average similarity $\overline{J(t, \gamma)}$ decreases in time: power-law fits of the form $\overline{J(t, \gamma)} =\sim \gamma^{\lambda(t)}$ yield $\lambda < 0$ for all t. The fit coefficient $\lambda(t)$ fluctuates around a typical value, because of seasonality effects, but does not changes substantially as a function of the starting time t (Fig. S5C), hence $\overline{J(t, \gamma)} = \overline{J(\gamma)}$. This implies that the rate at which the activity space evolves does not substantially depends on when the measure is initiated. We conclude that our data reflect the 'equilibrium' behavior of the monitored individuals. The fact that our dataset allow us

to replicate measures performed on other datasets obtained with different methods (see below) further confirms this finding.

3.1 Conservation of time allocation

Here we show how the results presented in the main text hold independently of how classes are defined. The individual sub-capacity are defined as $C_i(t)^{\Delta T} = |AS_i(t)^{\Delta T}|$. The average sub-capacities $\overline{C}^{\Delta T}(t)$ are constant in time for several choices of ΔT and different definitions of location. This is verified with the linear fit test as detailed in a previous section (Table S3). We consider both arbitrary time intervals and logarithmic spaced intervals. Individuals' net gain, with respect to changes of the AS subsets is zero. The study of the ratios $\sigma^{\Delta T}_{G,i}/\langle G^{\Delta T}_i\rangle$, where $\langle G^{\Delta T}_i\rangle$ is the average individual net gain for a class ΔT, and $\sigma^{\Delta T}_{G,i}$ is the corresponding standard deviation, yields that the ratio is higher than 1 for more than 98% of individuals, for all ΔT (Table S5).

3.1.1 Discrepancy relative to the randomized cases

Individual capacity is lower than it could be if individuals were only subject to time constraints. We showed this by randomizing individual temporal sequences of stop-locations for 100 times, and then comparing the average randomized capacity $\langle C_{rand,i}\rangle$ with the real capacity $\langle C_i\rangle$. We perform two types of randomizations (Fig. S8A):

- (1) Local randomization: For each individual i, we split her digital traces in segments of length 1 day. We shuffle days of each individual.
- (2) Global randomization: For each individual i, we split her digital traces in segments of length 1 day. We shuffle days of different individuals.

The individual randomized capacity $\langle C_{rand,i}\rangle$ averaged across time, (Figs. S8B and C), is higher than in the real case both for the global and the local randomization cases. We compute the KolmogorovSmirnov test-statistics (Table S7) to compare the real sample with the randomized samples. We reject the hypothesis that the two samples are extracted from the same distribution since $p < \alpha$ with $\alpha = 0.01$.

4 Additional measures

4.1 Establishment of individual activity space

At any point in time individuals allocate most of their time among few locations. For each user i, we consider the set of locations $AS_i(t) = \{\ell_1, \ell_2, ..., \ell_C\}$ seen in the W weeks preceding t at least twice and such that $T_{i,\ell}(t) > W \cdot 10$ min , where $T_{i,\ell}(t)$ is the total time of observation of location ℓ during the W weeks. We call this subset the activity space (AS). In Fig. S5A, we show that for $W = 10$ weeks, the AS contains typically $\sim 4\%$ of all locations seen during the same 10 weeks. Yet the time spent in these locations is on average $\sim 92\%$ of the total time.

Individuals are continually discovering locations, but only some among them join the AS. The probability that a newly discovered location at time t will be introduced in the activity space is $P_i(t) =$

$L_{i,AS}(t)/L_i$, where L_i are all the locations discovered by an individual i at t and $L_{i,AS}$ are the one that will be part of the individual's activity space. The average value of P across time. P stabilizes at $\overline{P} = 20\%$.

4.2 Evolution and composition of the activity space subsets

Individuals allocate time heterogeneously among locations, due to their different functions (homes, workplaces, shops, universities, leisure places...). We study time allocation between different classes of locations considering subsets of the activity space defined on the basis of the total visitation time. The subsets of the activity space $AS_i(t)^{\Delta T} \in AS_i(t)$ include all locations seen in the W weeks preceding t at least twice and such that $W * \Delta t(0) < T_{i,\ell}(t) < W * \Delta t(1)$ where $T_{i,\ell}(t)$ is the time of observation of location ℓ during the W weeks preceding t.

We test several choices of intervals ΔT. We find that when ΔT increases, the subsets are empty for many individuals, since no locations satisfy the above-mentioned criteria. In Fig. S9A-F, we show the distribution of average individual sub-capacities $\langle C_i^{\Delta T} \rangle$. Only subsets with small enough ΔT are significant for more than 50% of the population, and typically each individual has 1 location where he/she spend more than 48 hours per week.

The evolution of these subsets in time is quantified measuring the Jaccard similarity, as detailed for the entire AS in the main text. The average similarity $\overline{J(\gamma)}$ decreases in time for all ΔT. However, a power-law fit of the form $\overline{J(\gamma)} \sim \gamma^{\lambda}$, yields that the coefficients λ decreases as ΔT increases, since places where individuals spend more time are changed less frequently (Fig. S9G).

W	$\rho_{t,C}$	**p-value**	**Linear fit coeff** m	R^2	**PL fit coefficient** β
1	-0.057	0.58	-0.006±0.042	-0.049	-0.023±0.060
2	-0.035	0.735	-0.005±0.061	-0.026	-0.011±0.055
6	0.057	0.589	0.032±0.088	-0.015	0.024±0.061
10	0.107	0.325	0.046±0.103	-0.006	0.033±0.069
12	0.097	0.377	0.044±0.109	-0.011	0.033±0.073
16	0.078	0.49	0.038±0.121	-0.018	0.031±0.083
24	-0.172	0.145	-0.022±0.147	0.016	-0.015±0.108
32	-0.149	0.235	-0.023±0.176	0.014	-0.018±0.138

Table S1: **Conservation of the average capacity** For different windows sizes W the table shows the Pearson correlation coefficient, $\rho_{t,C}$, with the corresponding p-value testing the hypothesis $\rho_{t,C} = 0$ (there is no significant correlation between time and capacity at $\alpha = 0.05$ when p-value$> \alpha$); The linear fit coefficient m with the corresponding coefficient of determination R^2 (the coefficient of determination is negative because a horizontal line fits better than the best fit); The power-law (PL) fit coefficient β.

	locations, d=10	**APs**
1	-0.02±0.02	0.14±0.29
2	-0.02±0.03	0.18±0.45
6	-0.01±0.05	0.55±0.68
10	-0.01±0.06	0.73±0.81
12	-0.02±0.06	0.76±0.86
16	-0.02±0.07	0.76±0.95
24	-0.05±0.08	0.29±1.18
32	-0.05±0.10	0.24±1.45
52	-0.06±0.17	0.16±2.54

Table S2: **Conservation of capacity for different sliding window sizes and definitions of location**. The linear fit coefficient m with the relative error for several values of sliding window size W (rows) and definitions of location (columns).

	$\overline{m}$	$t-statistics$	$p-value$
locations, d=5m	-0.054 ± 0.048	-1.128	0.259
locations, d=10m	-0.073 ± 0.027	-2.677	0.008
APs	0.134 ± 0.370	0.360	0.718

Table S3: **Conservation of individual capacities.** The results of a t-statistics hypothesis test, testing the hypothesis $\overline{m} = 0$, where $\overline{m}$ is the sample average of individual linear fit coefficients. Results are shown for locations and APs (rows). The hypothesis $\overline{m} = 0$ is not rejected when p-value$> \alpha = 0.05$.

	1	2	6	10	12	16	24	32
locations,d=5m	0.99	0.98	0.97	0.98	0.98	0.97	0.95	0.94
locations,d=10m	0.99	0.98	0.97	0.98	0.98	0.97	0.96	0.95
APs	0.99	0.98	0.97	0.97	0.98	0.97	0.95	0.95

Table S4: **Individual gain is equal to zero** The percentage of individuals for which $|\langle G_i \rangle| < \sigma_i^G$ holds (the average individual gain is smaller than one standard deviation), for different definitions of location (rows) and time window sizes W (columns).

	10 min-1 h/w	1-6 h/w	6-12 h/w	12-24 h/w	24-48 h/w	48-168 h/w
locations,d=5m	0.98	0.98	0.96	0.92	0.86	0.76
locations,d=10m	0.98	0.97	0.96	0.9	0.79	0.66
APs	0.98	0.98	0.98	0.97	0.97	0.96

Table S5: **Net gain consistency with zero for different choices of ΔT.** The percentage of individuals for which holds $|\langle G_i^{\Delta T} \rangle| < \sigma_i^{G,\Delta T}$, where $|\langle G_i^{\Delta T} \rangle|$ is the average individual gain for a given category ΔT of locations, and $\sigma_i^{G,\Delta T}$ the corresponding standard deviation. The result is given for different definitions of location (rows) and several categories ΔT (columns) with a sliding window $W = 10$. The decrease of percentage for large ΔT is due to the fact that individuals have typically few places where they spend large amounts of time per week (typically just 1 for $\Delta T > 48$ hours (see Figure S9, hence the standard deviation and mean value are not meaningful in this case.

j / k	0	1	2	3	4	5	6	7
0		0.516	0.194	0.748	0.944	0.7	0.781	0.401
1			0.475	0.78	0.603	0.835	0.74	0.798
2				0.352	0.253	0.389	0.325	0.671
3					0.819	0.949	0.964	0.618
4						0.772	0.851	0.475
5							0.913	0.666
6								0.584
7								

Table S6: **Conservation of capacity.** The p-values, $P_{(j,k)}$, computed to test the hypothesis $H_{j,k}$: $\overline{C(\delta t_k)} = \overline{C(\delta t_j)}$ for locations with threshold $d = 5m$ (j is the row and k is the column, the matrix is symmetric). For all j, k, the hypothesis $H_{j,k}$ can not be rejected at $\alpha = 0.05$ since $P_{(j,k)} > n(j,k)\alpha/M$ (where $M = 28$ is the number of hypotheses and $n(j,k)$ is the rank of a pair (j,k) based on $P(j,k)$ in increasing order, [48]).

	KS statistics	**p-value**
local randomization	0.35	$8 \cdot 10^{-43}$
global randomization	0.96	$2 \cdot 10^{-321}$

Table S7: **Disagreement from the randomized series.** Results of the Kolmogorov-Smirnov test comparing the individual average capacities sample, with the randomized capacities sample. The hypothesis that the two samples are extracted from the same distribution is rejected at $\alpha = 0.05$ since $p - value < \alpha$, both for local and global randomization.

Δt	**Linear fit coeff**	R^2
Arbitrary time intervals		
0.17-1.0 hw	0.055± 0.076	0.006
1.0-6.0 hw	-0.023± 0.037	-0.034
6.0-12.0 hw	0.005± 0.009	0.142
12.0-24.0 hw	0.005± 0.006	0.376
24.0-48.0 hw	0.001± 0.005	-0.01
48.0-168 hw	0.003± 0.005	0.163
Logarithmic spaced intervals		
0.17-0.69 h/w	0.049±0.067	0.009
0.69-2.85 h/w	-0.015±0.039	-0.056
2.85-11.70 h/w	-0.002±0.019	-0.018
11.70-48.0 h/w	0.006±0.008	0.214
48.0-168 h/w	0.003±0.005	0.163

Table S8: **Conservation of time allocation.** The linear fit coefficient of the average capacity for different ΔT, and the coefficient of determination of the linear fit R^2.

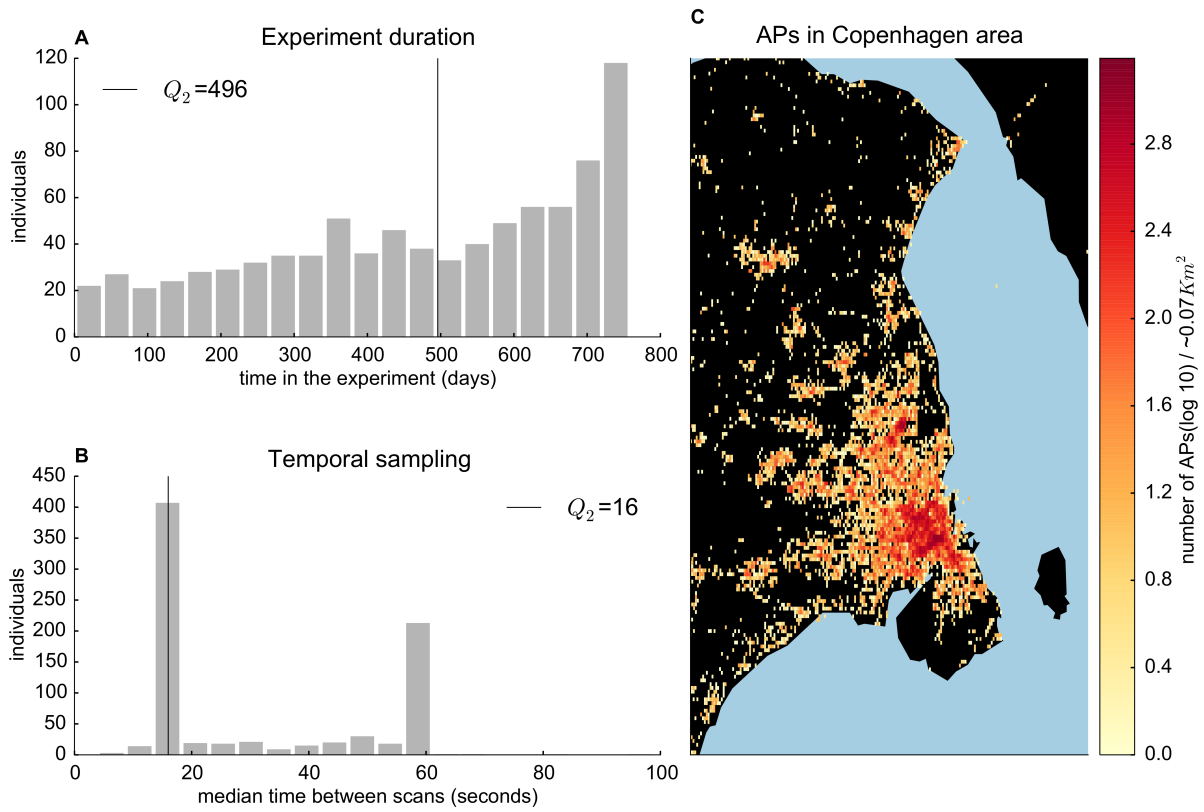

Figure S1: **Data description.** **A**) Frequency histogram of participants according to the total time in the CNS experiment. **B**) Frequency histogram of individuals according to the median time between consecutive scans. Q_2 is the median across the population. **C**) Heat map of the number (in log 10) of geo-localized APs in the Copenaghen area.

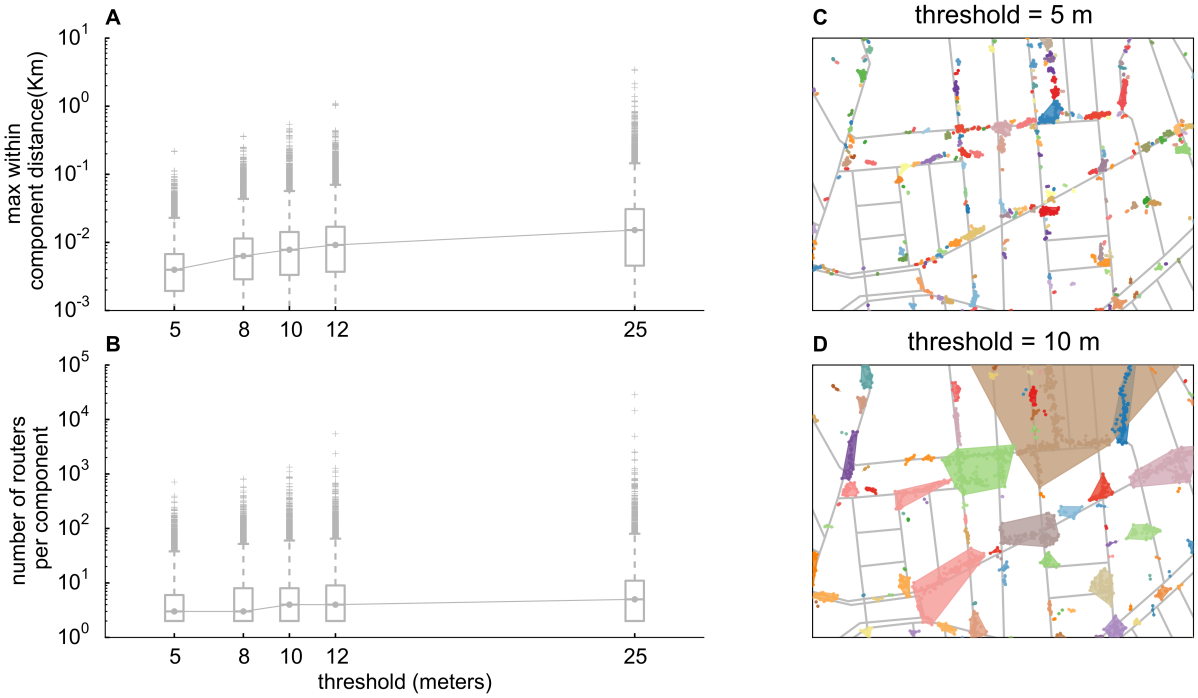

Figure S2: **From APs to "locations".** **A**) The boxplots of the maximal distance between pairs of geo-localized APs forming a location, as a function of the threshold d used to merge APs. Boxes are set at the 1st and 3rd quantile, while whiskers at 2.5% and 97.5%. **B**) The boxplots of the locations size (number of APs) as a function of the threshold d. **C-D**) An example of the clustering of APs located within Copenhagen city for thresholds $d = 5m$ (C) and $d = 10m$ (D). Dots corresponds to geo-localized APs, colored according to the location they belong to. Colored regions are the convex hulls of the set of APs in a same location. Grey lines are streets.

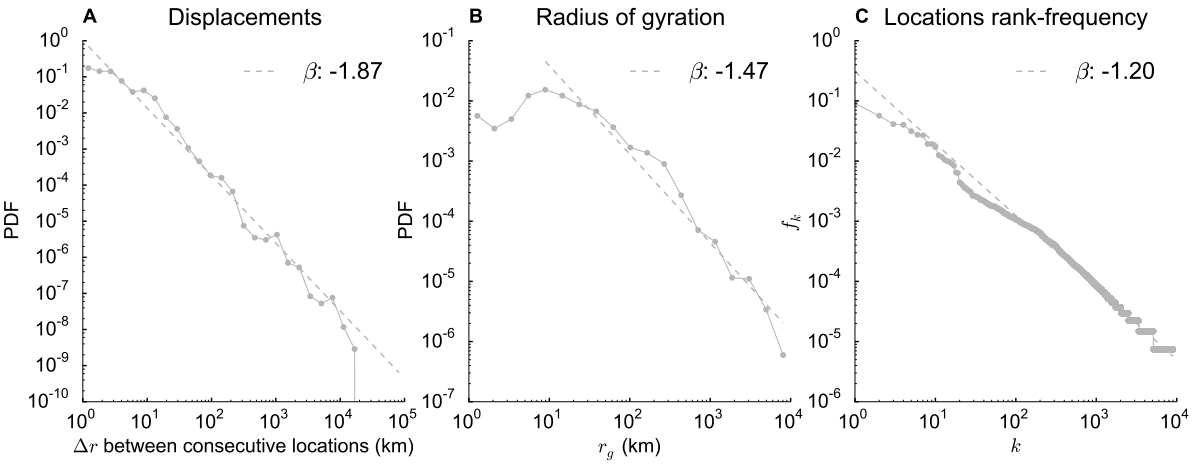

Figure S3: **Agreement with previous research.** **A**) The probability density distribution of jump lengths (in Km) between consecutive stop-locations (dotted line), and the corresponding power law fit with coefficient β (dashed line). **B**) The probability density distribution of individuals final radius of gyration $rg_i(t_{max})$, where t_{max} is the moment when an individual dropped the experiment (dotted line) and the corresponding power-law fit (dashed line) with coefficient β. **C**) The average visitation frequency f_k as a function of a location rank k across the population (dotted line) and the power law fit $f_k \propto k^{\beta}$ (dashed line).

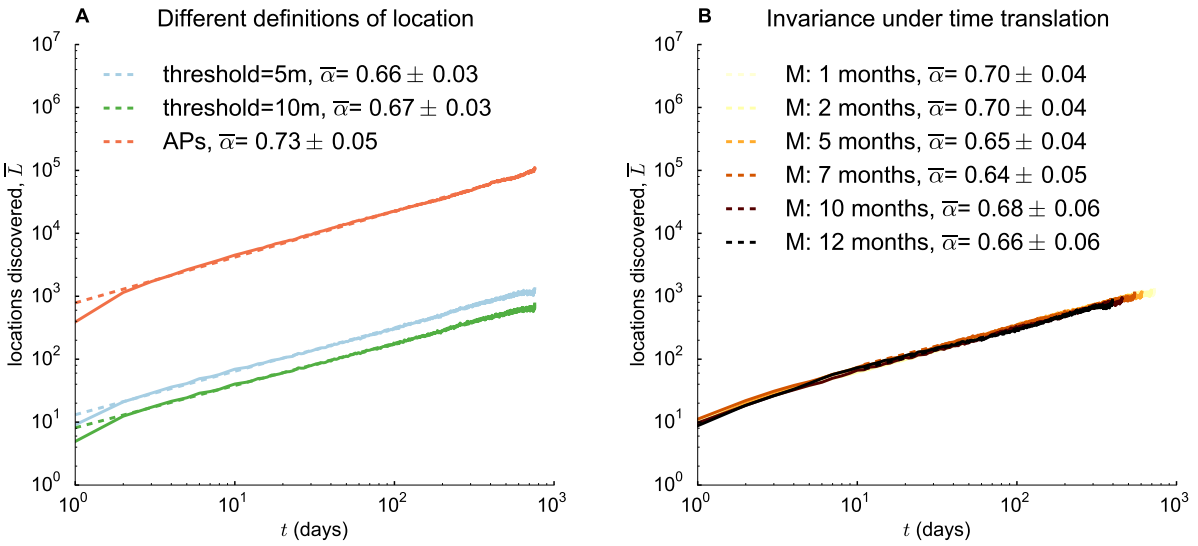

Figure S4: **Sublinear growth of individual locations. A**) The average number of locations discovered up to day t for different definitions of location, and the corresponding power-law fits (dotted line) with coefficient $\overline{\alpha}$. **B**) The average number of locations individually discovered in time, measured after waiting M months, and the corresponding power-law function fit with coefficients α (dotted lines) for different values of M.

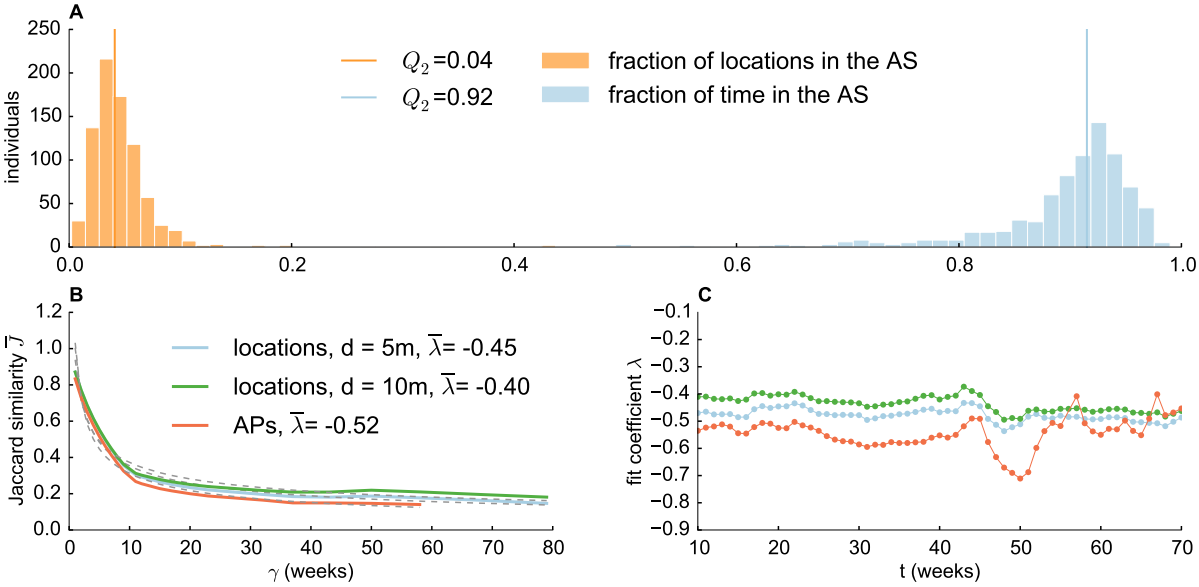

Figure S5: **Establishment and evolution of the AS. A**) Frequency histograms of individuals based on the fraction of all locations seen in a week that are part of the activity space (median across weeks, orange bars), and on the fraction of time of the week spent in the activity space (median across weeks, blue bars). The activity space is computed for $W = 10$ weeks, Q_2 is the median value across the population. **B**) The average overlap (Jaccard similarity) between the activity space at week t and week $t+\gamma$ (full line), and the corresponding power law fit $\overline{J(\gamma)} \sim \gamma^{\lambda}$ (dashed lines) with coefficient $\overline{\lambda}$ (dashed line) for different definitions of location. **C**) The PL fit coefficients $\lambda(t)$ as a function of the starting time of the measurement t, for different definitions of location.

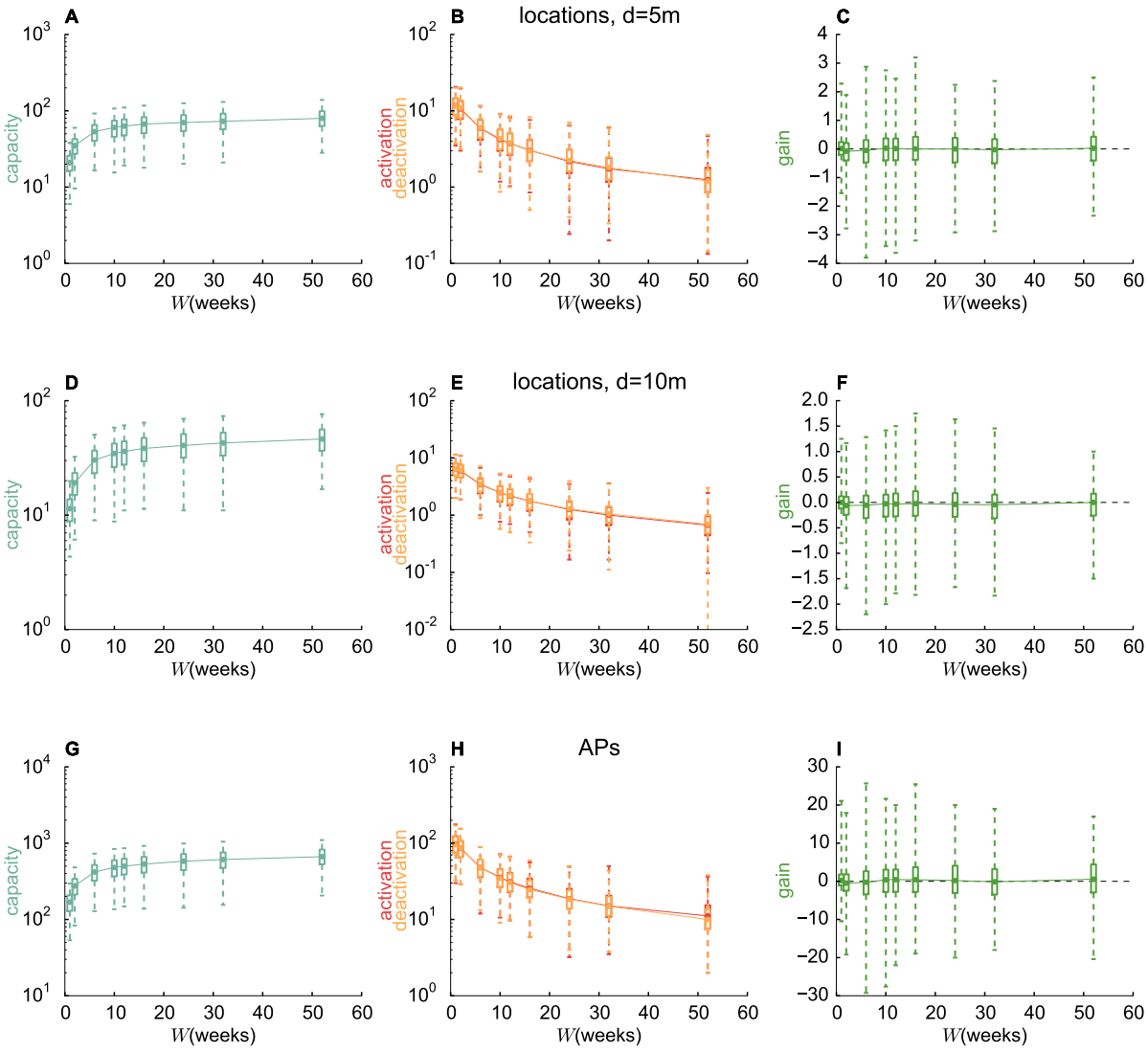

Figure S6: **Dependence on the window size for Capacity, Activation, Deactivation and Gain.** The boxplots of the individual average capacity $\langle C_i \rangle$ (left), individual activation $\langle A_i \rangle$ (center) and deactivation $\langle D_i \rangle$ (center) and gain $\langle G_i \rangle$ (right), as a function of the sliding window size for locations with $d = 5m$, $d = 10m$ and APs (from top to bottom). Boxes contains the population interquartile (25 to 75 percentiles) and whiskers contain the 95% of the population (2.5 to 97.5 percentiles).

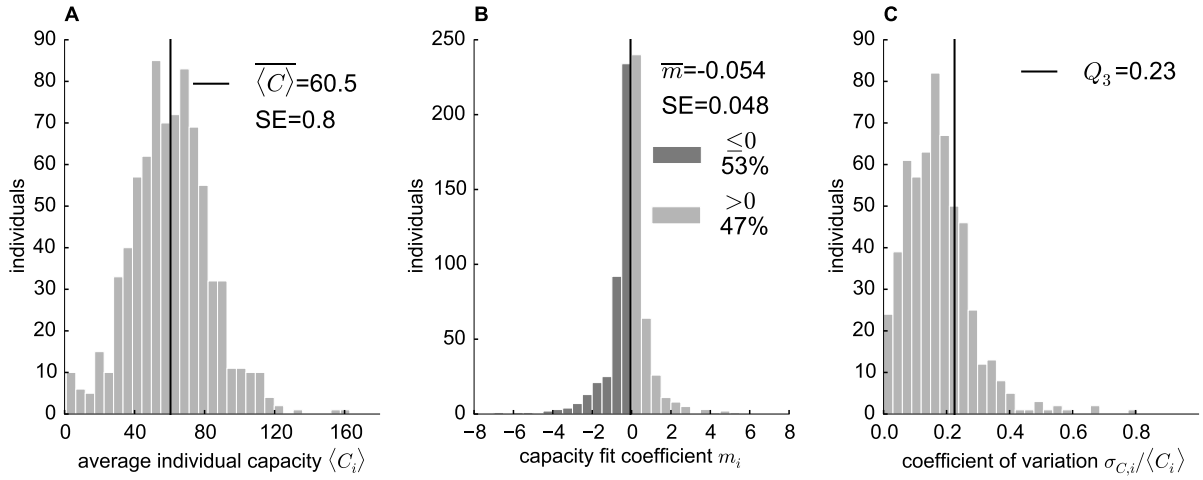

Figure S7: **Individual capacity: population homogeneity.** **A**) The frequency histogram of the average individual capacity $\langle C_i \rangle$ for locations with threshold $d = 5m$. The average value $\overline{C}$ (black line) has standard error SE. For APs, $\overline{C} = 484.3$ and $SE = 6.2$. For locations with threshold $d = 10m$, $\overline{C} = 34.5$ and $SE = 0.4$. **B**) The frequency histogram of individual fit coefficients m_i, where individual capacities $C_i(t)$ are modelled as $C_i(t) = a_i + m_i \cdot t$ for locations with thresholds $d = 5m$. The sample average coefficient $\overline{m}$, with corresponding standard error SE, is consistent with 0 under t-statistics hypothesis test at significance level $\alpha = 5\%$. The same results hold for locations with threshold $d = 10$, and APs. **C**) The frequency histogram of the coefficient of variation $\sigma_{C,i}/\langle C_i \rangle$, where $\langle C_i \rangle$ is the individual capacity averaged across time, and $\sigma_{C,i}$ the corresponding standard deviation for locations with threshold $d = 5m$ and weeks with time coverage higher than 80%. For locations with threshold $d = 10m$, $Q_3 = 0.22$, for APs, $Q_3 = 0.27$.

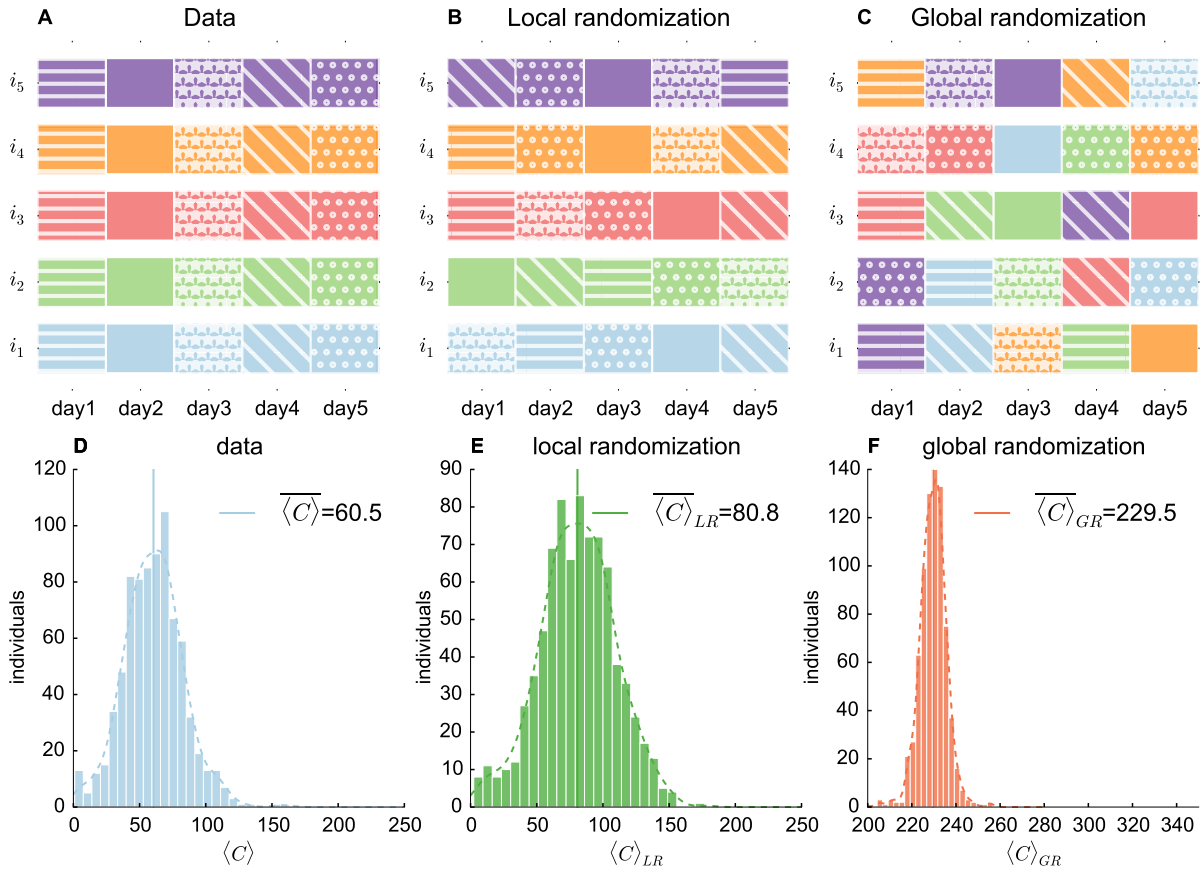

Figure S8: **Data randomization.** **A-B-C**) A schematic representation of local and global randomization. **A**) Individual time series for 5 individuals are divided into modules of 1 day length (each day has a specific color pattern). **B**) In the *local randomization* individual timeseries are shuffled preserving the module units. **C**) In the *global randomization* new sequences are created assembling together modules extracted randomly from the whole set of individual traces. **D**) The frequency histogram of the average individual capacity for data ($\langle C \rangle$), local ($\langle C_{LR} \rangle$) and global ($\langle C_{LR} \rangle$) randomizations, and the corresponding average values (full lines) computed across the population. Dashed line are gaussian kernel density estimations.The KolmogorovSmirnov test-statistics (Table S7) rejects the hypothesis that the three samples are extracted from the same distribution.

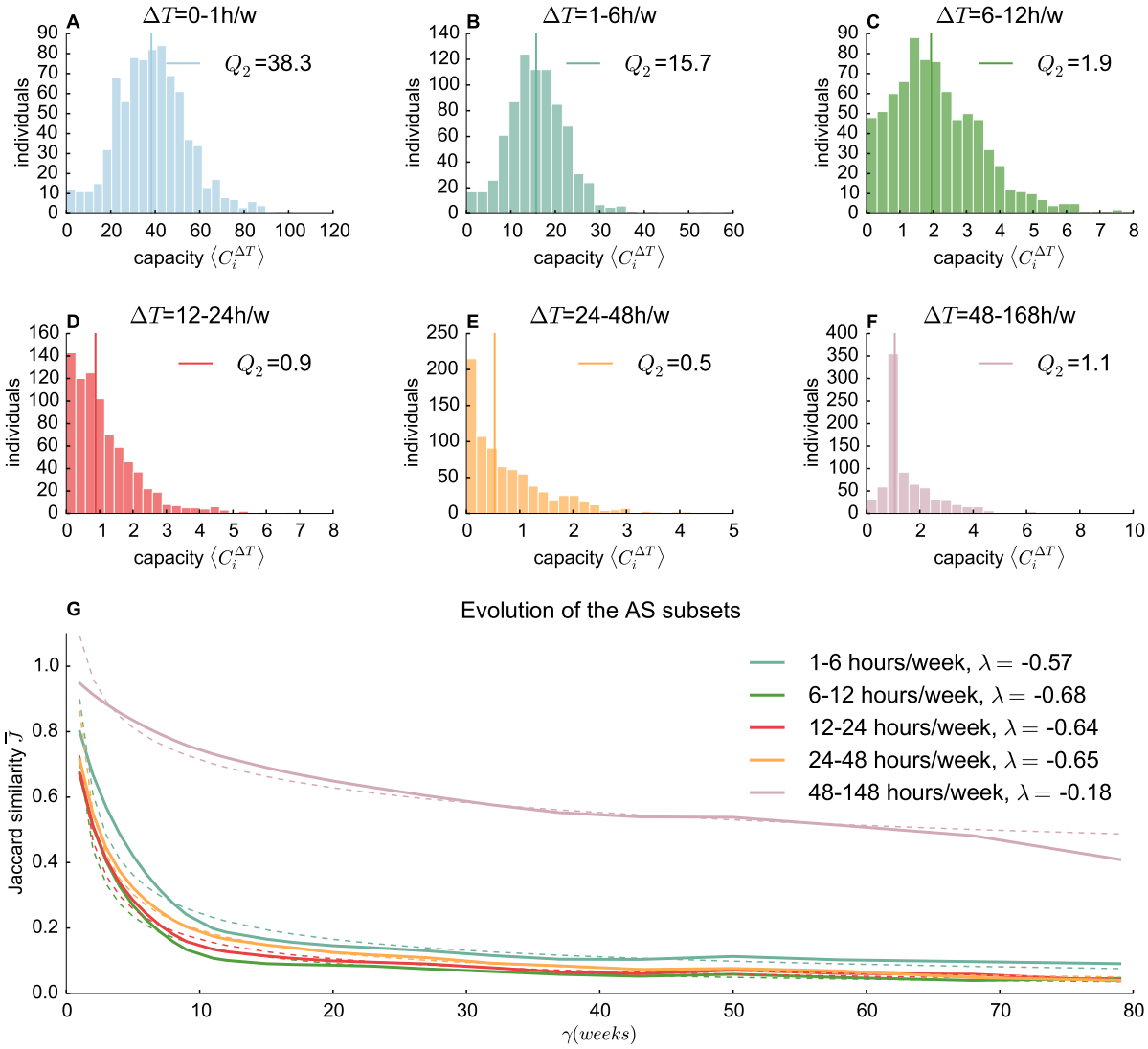

Figure S9: **Composition of the AS. A-F**) The distribution of the average individual capacity $\langle C_i \rangle^{\Delta T}\rangle$, considering locations seen for a time included in ΔT. **G**) The average overlap (Jaccard similarity) between subsets of the activity space at time t and $t+\gamma$ vs γ (thick lines), and the corresponding power-law fits $\overline{J(\gamma)} \sim \gamma^{\lambda}$ (dashed lines) with exponent λ, for different values of ΔT. Measures are performed for locations with thresholds $d = 5m$. Each line corresponds to a different choice of ΔT. It is worth noting that decay of the overlap for the > 48 class is slower than for the other classes, and that this overlap is, in general, higher. This implies that locations where individuals spend a great amount of time every week are more stable, in agreement with previous results on the stability of human mobility [1, 2, 3].

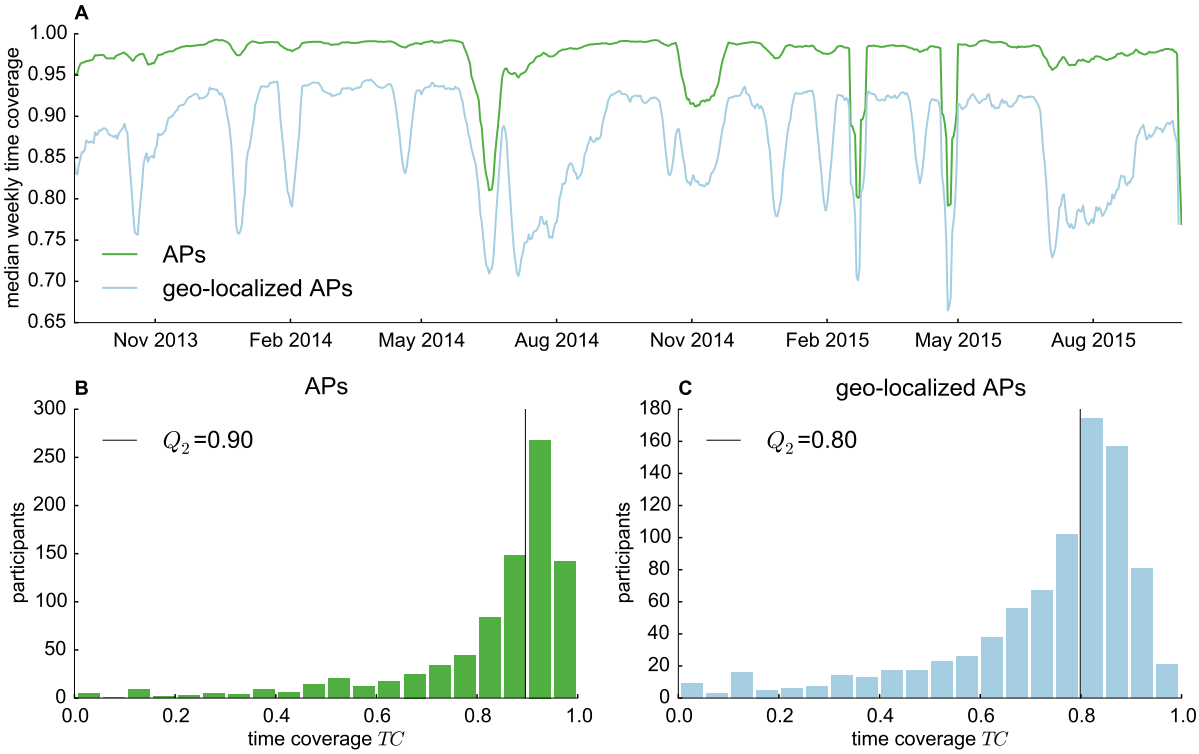

Figure S10: **High time coverage of the CNS WiFi dataset.** **A**) Median time coverage (fraction of time a user location is known) computed for all APs (green line) and geo-localized APs (light blue line). **B-C**) Frequency histogram of individuals based on their total time coverage, considering all APs (**B**), and only geo-localized APs (**C**). Q_2 is the median value across the population.