Mesoscopic Structure and Social Aspects of Human Mobility

James P. Bagrow1,2, Yu-Ru Lin3,4

1 Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, Illinois, United States of America, 2 Center for Complex Network Research, Northeastern University, Boston, Massachusetts, United States of America, 3 College of Computer and Information Science, Northeastern University, Boston, Massachusetts, United States of America, 4 Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts, United States of America

Abstract

The individual movements of large numbers of people are important in many contexts, from urban planning to disease spreading. Datasets that capture human mobility are now available and many interesting features have been discovered, including the ultra-slow spatial growth of individual mobility. However, the detailed substructures and spatiotemporal flows of mobility – the sets and sequences of visited locations – have not been well studied. We show that individual mobility is dominated by small groups of frequently visited, dynamically close locations, forming primary “habitats” capturing typical daily activity, along with subsidiary habitats representing additional travel. These habitats do not correspond to typical contexts such as home or work. The temporal evolution of mobility within habitats, which constitutes most motion, is universal across habitats and exhibits scaling patterns both distinct from all previous observations and unpredicted by current models. The delay to enter subsidiary habitats is a primary factor in the spatiotemporal growth of human travel. Interestingly, habitats correlate with non-mobility dynamics such as communication activity, implying that habitats may influence processes such as information spreading and revealing new connections between human mobility and social networks.

Introduction

Understanding human movement is essential for a range of society-wide technological problems and policy issues, from urban planning [1] and traffic forecasting [2], to the modeling and simulation of epidemics [3,4,5]. Recent studies on mobility patterns have shown that spatiotemporal traces are highly non-random [6,7,8], exhibiting distinct dynamics subject to geographic constraints [9,10,11,12,13,14]. Analytical models have been developed to reflect individual mobility dynamics such as the tendency to move back and forth between fixed locations on a regular basis [15]. When examining populations, movement patterns may be highly correlated with dynamics such as contact preference [9,11], yet this has not been well studied at the individual level. Previous work on human mobility has focused primarily on simple measures that forego the majority of the detailed information available in existing data. There is good reason for this, as basic approaches tend to be most fruitful for new problems. Yet these measures reduce an entire mobility pattern to a single scalar quantity, potentially missing important details and throwing away crucial information.

A number of approaches are available for studying the geographic substructure of individual mobility. One route is to perform spatial clustering [16] on the specific locations an individual visits, potentially revealing important, related groups of locations. However, such analysis is purely spatial, neglecting the detailed spatiotemporal trajectories available for each person, reducing their mobility to a collection of geographic points and ignoring any information regarding the flows, or frequencies of movement, between particular locations. At the same time, the raw spatial distance separating two locations may not be meaningful: a short walk and a short car trip typically cover very different distances in the same amount of time, and the cognitive and economic costs associated with air travel depend only mildly (if at all) upon distance [17]. Modeling frameworks such as the Theory of Intervening Opportunities [18] and the recently introduced Radiation model [19] further argue that raw distances are not necessarily the most effective determinant for travel. In this work we show the importance of incorporating how frequently an individual travels between two locations, which naturally accounts for spatial and dynamic effects while revealing the underlying spatiotemporal features of human mobility.

Results

Beginning from a country-wide mobile phone dataset [20,7,21,8,13,22,23,24], we extract 34 weeks of call activity for a sample population of approximately 90 thousand phone users.
Each call activity time series encodes the spatiotemporal trajectory of that user. (See Materials and Methods and File S1 for details about the data.) For each user we construct a directed, weighted mobility network capturing the detailed flows between individual locations (represented using cellular towers). Examples of both mobility networks and spatiotemporal mobility flows are shown in Figs. 1A and B, respectively. The recurrent and repetitive nature of human motion is clearly visible in Fig. 0B, where we explore the user trajectories vertically in time. We apply to each user’s mobility network an information-theoretic graph partitioning method known as Infomap [25], which uses the flows of random walkers to find groups of dynamically related nodes in directed, weighted networks. We do not use spatial or distance information in partitioning, instead Infomap mirrors the stochastic process underlying human mobility flows; see File S1 Sec. S3 for details. (Infomap’s underlying mechanism is further justified in this context by the results of [22].) The groups of locations that we discover, which we refer to as mobility “habitats,” will be shown to be crucial to both the spatiotemporal dynamics of human motion, and to the interplay between mobility and human interaction patterns. We rank habitats in decreasing order of phone activity, such that a user’s most frequently visited habitat is Habitat 1 or the primary habitat. We observe that human mobility is almost universally dominated by the primary habitat, where the majority of user call activity occurs—and thus it incorporates both home and work, home and school, or other major social contexts—along with a number of less active subsidiary habitats (see Fig. 1C, File S1 Fig. B, Sec. S3.2). We further see in Fig. 1D that most users possess 5–20 habitats, while only approximately 7% of users have a single habitat. Note that these habitats, unique for each member of the population, differ greatly from existing work on partitioning mobility or social connectivity [26,19,27], which instead focus entirely on partitioning a single geographic network aggregated from large populations.

**Spatial characteristics**

The spatial extent of a user’s total mobility pattern has been shown to be well summarized by a single scalar quantity, the radius of gyration, or gyradius, \( R_g = \sqrt{\langle r_i^2 \rangle} \), where \( r_i \) is the spatial position of phone call \( i \) and \( \langle \cdot \rangle \) is the user’s center of mass [7]. In addition to using the global gyradius we also compute the reduced radius of gyration \( r_g(h) \) for each habitat \( h \), considering only those locations and calls contained within that habitat. In Fig. 2A we plot the population distributions of the first three habitat’s \( r_g \), compared with the total gyradius \( R_g \) considering all places called from all visited locations. This shows that the spatial extent of habitats tends to be far smaller than the total mobility, often by an order of magnitude, and that most users have a habitat \( r_g \) between 1–10 km. See also File S1 Fig. D. In Fig. 2B we study the functional dependence of the primary habitat’s gyradius, \( r_g(h_1) \), versus \( R_g \). We uncover an intriguing power law scaling relation characterized by two regimes, where \( r_g(h_1) \sim R_g^\alpha \) with \( \alpha = 1 \) for \( R_g < R_s \), \( \approx 5 \) km, and \( \alpha = 1/3 \) for \( R_g > R_s \). The linear relationship below this critical radius \( R_s \) indicates that those users (roughly 8% of the population) are mostly characterized by a single habitat. (In fact, only 54.8% of users with \( R_g < 5 \) km have one habitat, but that 97.6% of their calls on average occur within their primary habitat.) But once a user’s range extends beyond this critical 5 km cutoff (true for 92% of the population) a new regime emerges where multiple habitats exist and tend to be far smaller and more spatially cohesive than the total mobility (since \( \alpha < 1 \)). (For users with \( R_g > 5 \) km, only 2.9% have one habitat and the primary habitat accounts for 78.7% of activity on average.) Finally, in Fig. 2C we show the geographic distance \( d(h_1,h_2) \).

---

**Figure 1. Habitats reveal the spatiotemporal substructure of human mobility patterns.** (A) Spatial trajectories of two users, one traveling to a large number of locations and another covering a smaller range. Node size indicates the amount of time spent at a particular location (as quantified by mobile phone activity), node color represents the location’s habitat detected using Infomap (see Methods), and line width approximates the number of trips between locations. Habitats are ordered by call volume such that Habitat 1 contains the most calls. (B) Exploding the spatial trajectories from A in time (vertical axis), the recurrent nature of human mobility becomes evident, with a number of trips featuring both consistent destinations and consistently repetitive occurrence (zoom). These features are the root cause of the high predictability that human motion is known to possess. (C) The daily call dynamics of the three most active habitats, as well as the overall dynamics (summed over all habitats). The primary habitat contains the majority of temporal activity. We see that User 1 tends to occupy his or her second and third habitats primarily at night, while User 2 is more evenly distributed. (D) The distribution of the number of habitats per user. The median number of habitats is 11. Due to their typical heterogeneity, we characterize population distributions using percentiles, proportional to the cumulative distribution.

doi:10.1371/journal.pone.0037676.g001
between the centers of mass of the two most heavily occupied habitats, as a function of $R_g$. This also exhibits a power law scaling, $d(h_1, h_2) \sim R_g^\beta$ with $\beta=0.81\pm0.02$. These distances tend to be far larger than the total $R_g$ (gray line), indicating that the magnitude of $R_g$ is primarily determined by movement between spatially cohesive and well separated habitats.

**Temporal characteristics**

Given the importance of habitats to the spatial extent of human motion, one must ask: how do these habitats form and evolve over time? To what extent are the temporal dynamics of human movement reflected in the evolution of these habitats? Recently, considerable effort has been undertaken to understand the intriguing temporal features of human mobility, including the previously observed ultra-slow growth in time $t$ of $R_g(\log t)^\gamma$, with $\gamma>1$ [7,15]. Given the contribution of habitats to the magnitude of $R_g$, shown in Fig. 2, a primary question becomes: how do habitats impact these temporal features? For example, how do individual habitat $r_g$’s evolve over time, compared with that of the total $R_g$?

In Fig. 3 we study the temporal evolution of $r_g$ and $R_g$ by considering only those calls occurring up to time $t$, from either individual habitats or all locations, where $t=0$ is the time of the user’s first call. In Fig. 3A we plot the time series of $r_g(h_1)$ and $R_g$, normalized by the final values of each respective series. We observe that $r_g$ saturates at its final value more quickly than the total mobility’s $R_g$. To further quantify this saturation, we plot in Fig. 3B the ratio between $r_g(h_1)$ and $R_g$ as a function of time, for groups of users with different final values of $R_g$. We observe increasingly rapid saturation of $r_g$ as the total $R_g$ increases. This implies that the primary habitat is explored more quickly than the total extent of a user’s mobility pattern and that users who cover large distances explore their primary habitat more quickly relative to their total mobility than users who traverse relatively smaller regions. This is particularly interesting as one may initially expect such exploration to be at a constant rate relative to their total $R_g$.

In Fig. 3C we study the temporal evolution of $r_g(t)$ for the first three habitats, averaged over users with $R_g \approx 30$ km. We observe approximately logarithmic growth, $r_g(h_1) \sim \log t$, for the primary habitat (slower growth than that observed in [7,15]) while subsidiary habitats’ gyradii $\sim (\log t)\delta$, with $\delta>1$ [growth more similar to [7,15]]. However, this analysis neglects an important detail: users do not begin exploring all of their habitats at the same time. Therefore in Fig. 3D we plot the same population-averaged radii, but now individually shift each user’s time series of $r_g(h)$ by a time $t_0(h)$, the time the user first entered habitat $h$, not simply made his or her first global call. Doing so accounts for the waiting times for users to visit habitats within our observation window. With this crucial correction we reveal for all habitats purely logarithmic growth in $r_g$, implying a universality in the exploration of habitats (which differ only in their overall spatial scale, with the primary habitat tending to be the most compact). Thus, the polylogarithmic growth of $R_g$, where $R_g$ is initially small then grows faster than logarithmic in time, is primarily due to the temporal delay it takes users to exit their respective primary habitats and then rapidly traverse a relatively large distance to reach their other habitats. We further study these habitat entrance times in File S1, Sec. S3.2 and Fig. E.

Social characteristics

Finally, a major question in the realm of mobility and human dynamics is the connection between spatiotemporal dynamics and other activity patterns [9]. For example, information spreading in heterogeneous systems of agents is a process that involves both the spatiotemporal mobility of the agents and their long-range communication activities. In this context, would the currently occupied habitat affect or be affected by how a user chooses a particular communication partner to engage? Such questions can also be addressed with mobile phone data, where phone communications capture a primary mode of information diffusion on the underlying social network [20]. To begin, we first recall a result from González et al. [7]. They found that users occupy locations following a Zipf law, where the probability $Pr(L)$ to visit the $L$-th most frequented location follows $Pr(L) \sim L^{-1.2}$. We reproduce this result in Fig. 4A. Interestingly, we discover (Fig. 4B) a potentially identical mechanism for how users choose to contact their communication partners, i.e. the probability $Pr(C)$ for a user to call his or her C-th most contacted partner also follows $Pr(C) \sim C^{-1.2}$. See also [28]. Finally, a number of users within our population have contacts that are also within the population, meaning we have habitat information for both users. An interesting question is: how similar are the habitats of users in

---

**Figure 2. Spatial properties of mobility habitats.** We characterize each habitat’s spatial extent by computing the radius of gyration $r_g(h)$ considering only calls placed from locations within habitat $h$. (A) The distribution of habitat radii over the population shows that the primary habitat tends to be more spatially compact than the less frequented habitats, though most are consistently smaller than the total $R_g$ computed using all phone activity. (B) The growth in the radius of the primary habitat $r_g(h_1)$ as a function of total radius $R_g$. For $R_g < R_c \approx 3$ km, we see $r_g(h_1) \approx R_c$, indicating that those users are characterized by a single habitat. In contrast, $r_g(h_1) \sim R_g^{1/3}$ for $R_g > R_c$. Since approximately 92% of the population have $R_g > 5$ km, the majority of users exist in a regime where their primary habitat encompasses a potentially far smaller spatial region than their total mobility. (C) For users with multiple habitats, the distance $d(h_1, h_2)$ between the first and second habitat’s centers of mass is consistently greater than $R_g$ (gray line) and exhibits power law scaling, $d(h_1, h_2) \sim R_g^{\beta}$, with $\beta=0.81 \pm 0.02$. Taken together, we see that most habitats are both well separated and spatially compact, and that the magnitude of $R_g$ is primarily due to movement between these habitats.

doi:10.1371/journal.pone.0037676.g002
close communication, and will this similarity be lower for pairs with less frequent interaction? We measure the similarity between the primary habitats of pairs of users interacting with one another by computing the relative number of locations the habitats have in common (see Methods and materials). In Fig. 4C we plot this Habitat similarity as a function of contact rank C. We see that, despite the Zipf law in Fig. 4B, users’ primary habitats tend to be highly similar to the primary habitats of their most contacted ties. Nevertheless, there is little dependence on contact rank: one partner that is contacted an order of magnitude less often than another has almost the same primary habitat similarity. In other words, it takes very little communication to generate considerable habitat overlap [10]. Meanwhile, control habitats, generated by randomly distributing each user’s visited locations between their contacts (see Methods and materials) show smaller similarity and no effective trend.

We further characterize the “interaction concentration” of a user by introducing PMFC, the probability that the next call placed by the user goes to that user’s Most Frequent Contact, the partner that is most often in contact with the user. Users with a small PMFC tend to distribute their calling activity more evenly across their partners, while users with large PMFC are more concentrated and focus much of their attention upon a single individual. In Fig. 5 we study how PMFC depends on the properties of a user’s mobility pattern. First, in Fig. 5A we show the distribution of PMFC over the user population. Most users possess 0.1 ≤ PMFC ≤ 0.4 while very few users have either very small or very large PMFC. In Fig. 2B we connect this interaction concentration with the user mobility patterns by showing that the mean PMFC decays as the number of habitats a user visits grows. This means that users who travel broadly, leading to complex mobility patterns and multiple habitats, tend to distribute their communication activity more uniformly over their contacts. Next, in Fig. 5C we quantify how PMFC varies with the total gyradius Rg. We see a surprising connection to a previous result: For users with small Rg, the PMFC is small but grows as Rg grows. This continues until Rg ≈ Rf, the same critical radius that appeared in Fig. 2B. Above Rf, we see that PMFC now slowly decays with Rg. To further understand this, we plot in Fig. 5D the fraction of reciprocated contacts freciprocal (see Materials and methods) as a function of Rg. The plot exhibits a roughly similar trend as Fig. 5C: freciprocal grows while Rg < Rf, then, above the same critical radius, freciprocal decays slowly with Rg. This decay relative to the peak value at Rg = Rf, is slower for freciprocal than for PMFC.

Taken together, Figs. 5C and D show that when Rf > Rg, user communication activity—both how much they concentrate upon their MFCs and how many of their ties are reciprocated—depends only weakly on Rg. However, those users with low Rg tend to show distinctly different behavior, both being less concentrated on their MFCs compared with most users, and making a larger number of non-reciprocated contacts (File S1 Fig. C). Since users with Rg < Rf, primarily possess a single habitat, these results imply that the mechanism governing how users distribute their activity over their contacted partners may differ for those users with a single
habit compared with those users whose mobility leads to multiple habitats. We used Kendall rank correlation and associated hypothesis tests [29] to verify the statistical validity of the observed relationships. See File S1 Sec. S6 and Table A for hypothesis tests between these and additional measures.

The mobile phone data also provides demographic information for the majority of users, specifically self-reported age and gender. In File S1 Sec. S4 we study the results of Fig. 5 after decomposing the sample into age and gender groups. One may expect these results to change when focusing on these different subgroups. Yet in Figs. F and G we find qualitatively similar results to Fig. 5 with only small differences: \( P_{MFC} \) tends to be slightly higher for women than for men, and increases with user age. After considering these demographic dependencies on \( P_{MFC} \), we observe the same relationships between communication activity and mobility.

Discussion

We have shown that accurately understanding human mobility requires an analysis using the complete spatiotemporal flows captured for each user. Basic measures such as the gyroid radius \( R_g \) constitute an excellent starting point, but such single scalar quantities simply cannot capture the full complexity of an individual mobility pattern. As the quality and quantity of available data increases, we expect our understanding of the various factors shaping human mobility to continue to improve.

Given that users spend the majority of time occupying their primary habitats, understanding the detailed features of the primary habitat will be crucial for applications such as search and rescue during emergencies [23] or containing the spread of epidemic diseases [3,4,5], since most users will be within their primary habitats at the onset of such events. Meanwhile, detailed information regarding unusual trips away from the primary habitat may prove useful both for curtailting diseases and for optimizing transportation infrastructure and energy usage. Likewise, the universal logarithmic scaling laws for intra-habitat mobility uncovered in Fig. 3D are not accounted for by current modeling frameworks [15]; more effort may be necessary to acceptably model the microscopic structure of individual human motion. The connections we reveal between communication dynamics and human mobility may have important consequences for understanding the spread of information or rumors through a population, as such processes may spread both spatially and socially [30]. Further investigation of such connections may prove fruitful in a number of areas, including information diffusion and social contagion.

Materials and Methods

Dataset

We use a large-scale, de-identified mobile phone dataset, previously studied in [20,7,21,8,15,22,23]. We sample approximately 90 thousand users from the total dataset, according to the activity criteria introduced in [8] (see also File S1 Fig. H). We sample nine months of phone activity for each user. A “call” is either a text message or a voice call, and we use the cellular tower that handled the communication to represent the location \( L(t) \) of a call made at time \( t \). Call times are kept at an hourly resolution. The coordinates of these towers are used to compute the radii of gyration [7]. Phone call recipients determine the communication partners of a user. Since a single phone call between two individuals may not represent a meaningful tie, we consider user \( B \) to be a partner of user \( A \) only if we observe at least one reciprocated pair of calls (\( A \) called \( B \) and \( B \) called \( A \)) [20]. We do not require user \( B \) to be in our sample population, except when we compute habitat similarity. We define the fraction of reciprocal ties for user \( A \) as

\[
\text{f}_{\text{recip}}(A) = \frac{\sum_A X(A,B) X(B,A)}{\sum_A \sum_B X(A,B)}
\]

where \( X(A,B) = 1 \) if \( A \) contacted \( B \) at least once, and zero otherwise.

Finding mobility habitats

For each user we convert their trajectory \( Z = \{L(t_1), L(t_2), \ldots, L(t_f)\} \), with \( t_i > t_j \) for \( i > j \), into a weighted digraph where the weight on link \( L_i \rightarrow L_j \) represents the number of times the ordered pair of locations \( (L_i, L_j) \) was observed in \( Z \) (File S1 Fig. A). The community discovery method Infomap [25] was applied to each digraph, using the default parameters (10 attempts and self-loops ignored). The discovered groups of locations are the habitats for that user. Habitats are ranked by total number of calls.

Habitat similarity

For a user \( A \) with contact \( B \), both present in our sample, we define the similarity \( S(A,B) \) between their primary habitats to be the Jaccard coefficient between the sets of locations comprising those habitats. If these sets are disjoint \( S(A,B) = 0 \), whereas \( S(A,B) = 1 \) if they overlap completely.
Controls
It is important to understand the significance of the results we have presented here, in particular whether the results associated with the habitats we find are meaningful. We compute null or control habitats, generated for each user, by randomly assigning that user’s visited locations to habitats while preserving the number of habitats and the number of locations within each habitat. This strictly controls for the habitat size distributions while testing the effects of habitat membership. In File S1 Fig. 1 we further show that the pure logarithmic time evolution is absent in control habitats, indicating that the temporal evolution we observe in Fig. 3D is not due to the relative sizes (numbers of locations) of the habitats, nor to simply the number of habitats, but due more fundamentally to their spatial structure and the spatiotemporal flows of the users. In Fig. 1 we see that these control habitats have lower similarity than the actual habitats. See File S1 Sec. S6 for details.

References

Supporting Information
File S1 Supplementary text and figures. (PDF)

Acknowledgments
We thank F. Simini, J. Menche, J.-P. Onnela, S. Lehmann, W.-T. Chung, D. Brockmann, B. Uzzi, and D. Lazer for many useful discussions and A.-L. Barabási, D. Lazer, B. Uzzi and D. Brockmann for support. We gratefully acknowledge the hospitality of the LazerLab at Northeastern University, supported by MURI grant #G00003585. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding source.

Author Contributions
Conceived and designed the experiments: JPB YRL. Performed the experiments: JPB YRL. Analyzed the data: JPB YRL. Wrote the paper: JPB YRL.