Electronic Supplementary Materials: Coupling Human Mobility and Social Ties

Jameson L. Toole*

Engineering Systems Division, MIT, Cambridge, MA, 02144

Carlos Herrera-Yaqüe

Department of Applied Math, Universidad Politécnica, Madrid, 28040, Spain

Christian M. Schneider and Marta C. González

Department of Civil and Environmental Engineering, MIT, Cambridge, MA, 02144

SUPPORTING ONLINE MATERIALS

Data

Our data consist of anonymized call detail records collected from three cities (R1,R2, and R3) in two different industrialized countries. The same provider was used for the two cities in the same country (R1 and R2), while another provided the remaining city (R3). In R1 and R2, data cover 15 months while R3 contains 5 months of data. In total there are over 1 billion events contain the time and duration of a communication event between a caller and callee as well as the towers used by one (in the case of R3) or both (in the case of R1 and R2) of the users. Though data sharing agreements to protect privacy prevent us from sharing the locations of each region, they are major metropolitan areas with densities closely matching that of Boston, MA, USA.

Social Network Extraction

To build the social networks for each city, we employ the following procedure. First, we consider only users that appear in over 200 communication events within each city's metro region over the course of the entire data collection period. Second, we only draw an edge between two users if they make more than two calls between them during that time. Properties of the three networks as well as the number locations (cell towers) within each metro region can be found in Table S1.

TABLE I. Basic statistics on the networks and spatial extent of each region considered.

City	Nodes	Edges	$\langle k \rangle$	Towers
R1	133,587	997,287	14.9	249
R2	183,486	2,487,661	27.1	447
R3	635,731	4,197,093	13.2	935

Metric Definitions

Cosine Similarity, $\cos\theta$**:** In this work the cosine similarity is defined as the cosine of the angle between the location vectors of two users, $\cos\theta_{i,j} = \frac{\mathbf{v_i}\cdot\mathbf{v_j}}{|\mathbf{v_i}||\mathbf{v_j}|}$. In general, cosine similarity can take values from -1 to 1, but in our case it is non-negative as it is impossible for location vectors to have negative elements. This restricts angles to between 0 and $\frac{\pi}{2}$. We use the cosine similarity as a measure of how similar visitation patterns are between two users.

predictability, $\frac{|\hat{\mathbf{v}}|}{|\mathbf{v}|}$: Predictability provides an upper bound on how much of a target user i's visitation patterns can be reconstructed by the visitation patterns of a set of other users, F. In general, F can be made up of any users, but here we define it as the set of social contacts called by user i. Location vectors exist in an L-dimensional location space, where L is the number of unique locations that can be visited. In practice, most users visit only a small number of possible locations and the locations vectors of the users in F typically span only a subspace of the entire location space. If this subspace contains most of user i's location vector, then we can reconstruct their mobility patterns with a high degree of accuracy. If, however, user i's location vector is orthogonal to each vector in a basis of this subspace, then none of the user's visits can be recovered.

To quantify this, we first construct a $|F| \times L$ matrix $\mathbf{A} = [v_{f_1}, v_{f_2}, \ldots, v_{f_n}]^\intercal$ where f_j are contacts in F. We next use qr-decomposition to construct an orthonormal basis $B = q_1, \ldots, q_{|F|}$ of $\mathbf{A}$ that spans the subspace of the entire L-location space that is defined by the users in F. We can then project the original location vector of user i into this subspace to create the best approximation $\hat{\mathbf{v}}$ of that user's visitation patterns that can be constructed by users in F: $\hat{\mathbf{v}} = \sum_{j=1}^{|F|} \langle q_j, \mathbf{v} \rangle q_j$. To measure the accuracy of this approximation, we take the ratio it's magnitude to the magnitude of the original location vector for the target user, *predictability* $= \frac{|\hat{\mathbf{v}}|}{|\mathbf{v}|}$. Predictability can take values between 0 and 1 where the former indications none of the user's visits can be reconstructed by linear combination of users in F and 1 indicates all of a user's visits can be recovered.

We caution, however, against interpreting intermediate values of predictability as "fraction of visits correctly predicted". The magnitude of vectors are computed using the L2 norm are thus not equivalent to comparing percentages of visits recovered. The two values, however, are highly correlated in this case as there can be no negative elements in location vectors or their approximations. Finally, we note that predictability here is an upper bound approximations to a user's location vector using a linear combination of their contacts visits. Computing predictability requires full knowledge of a user's location vector. In the absence of this information, some proxy must be identified to replace the coefficient $\langle q_j, \mathbf{v} \rangle$. We encourage future work exploring this.

Number of Unique Locations Visited, S: The number of unique locations visited by a user is denoted as S. It can be computed directly from a user's location vector as $S = \sum_i sign(v_i)$.

Degree, k: A user's degree in the social network is denoted as k.

Contact Rank, r: For each user, we assign a contact rank r as a measure of tie strength to every neighbor in a user's ego network. We rank every contact of a user according to the number of calls made between them and assign a value of $r = 1$ to the contact exchanging the highest number of calls with the ego. We note that contact rank is not necessarily symmetric. User i may rank as user j's 3rd most called contact while user j is only user i's 10th most called contact.

Jaccard Similarity: The Jaccard similarity is a measure of set intersection. It is computed as the fraction of elements from set A that exist in set B. In the context of our social network, it is a measure of tie strength is defined as the fraction of contacts shared by two users i and j: $jaccard(i, j) = \frac{F_i \cap F_j}{|F_i||F_j|}$, where F_i is the set of neighbors contacts of user i.

Entropy, $H(f^{calls}_{i \to j \in C})$: Entropy is traditionally used as a measure of randomness or disorder. In the context of this work, we measure the information entropy of various probability distributions. Distributions with probability mass spread more evenly across all possible outcomes are considered to be "more random" and have higher entropy than distributions where a single outcome is far more likely than all others. Shannon's information entropy for a random variable X is computed as $H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$. In our case, we measure the entropy of the distribution of call frequency of a user to his or her social contacts. High values of entropy are calculated when a user distributes their calls evenly to all social contacts while low values are observed when they call a small number of contacts far more than the rest.

Controlling for number of calls

While mobile phones make excellent passive sensors of social behavior and mobility, they suffer from non-uniform sampling rates. Information is only recorded when a user uses the device leaving more observations at certain times of the day or week than others. Moreover, different users may use their devices more or less depending on habits or socio-economic variables. Because of this, we are careful to ensure that any metrics we measure in the data are not biased by different sampling rates.

Figure S1 shows the distributions of four metrics in each region for groups of users with similar numbers of calls over the observation period. In general, we find that calling frequency of users does not affect these distributions with the exception of the number of unique locations visited S, which increases with calling frequency. However, even in these cases the shape and trend of the distribution remains the same for each group with only the means shifted. Finally, we note that for region R3, the number of unique locations visited takes on a slightly different shape than regions R1 and R2 due to the fact that we only obtain location information for callers in this city and not for receivers as well. Our new metrics of mobility, cosine similarity and predictability, are least affected by different sampling rates.

We perform the same analysis for correlations between social behavior and mobility. For users with a given number of calls, we correlate their social metrics such as degree or the entropy with which they distribute calls to contacts with mobility metrics. We find that, similar to the distributions, most of these correlations do not depend on the number of calls made by a user. In cases where there is dependence, the trends hold within groups of users that make the same number of calls (Figure S2).

Controlling for Degree

We measure the entropy of the distribution of calls that each user makes to his or her contacts. Users with higher entropy spread their calls evenly amongst social ties, while lower entropy means most calls go to fewer. The degree of a node sets an upperbound on the entropy a user can have. Thus, any correlations we measure may be biased by differences in the degrees of each user. To control for this, we plot correlations of call entropy to other metrics for groups of users with the same degree. Figure S3 shows that these trends are unaffected by differences in degree.

Social Distance and Geographic Similarity

We compute the average cosine similarity between two nodes separated by a social distance of k hops. Much like previous studies of homophily within social networks, we find that geographic similarity is elevated for two individuals who call each other, but this increase in similarity extends outward up to three hops away after which users are as similar as they would be to random users (Figure S4).

Clustering

The k-means clustering algorithm must be seeded with the number of clusters to find a-priori. In order to identify a reasonable number of clusters, we run the algorithm for multiple values of k and examine the resulting clusters as well as the silhouette score for each choice. The silhouette score decreases as the number of clusters increases indicating that there is little added benefit from additional splitting (Figure S5). Moreover, when examining the centroids of clustering results, the three main clusters identified break into similar groups that show small differences such as on weekends or in absolute similarity level (Figure S6). To ensure our results are not an artifact of the clustering algorithm chosen, we also perform clustering using a hierarchical, agglomerative clustering technique using Ward linkage. In each region, we obtain clusters that match those found with k-means very closely (Figure S7).

Ego-network link type mixes

To ensure that our results are not an artifact of call frequency we plot the mix of an individual's social network as a function of the number of calls they make (Figure S8).

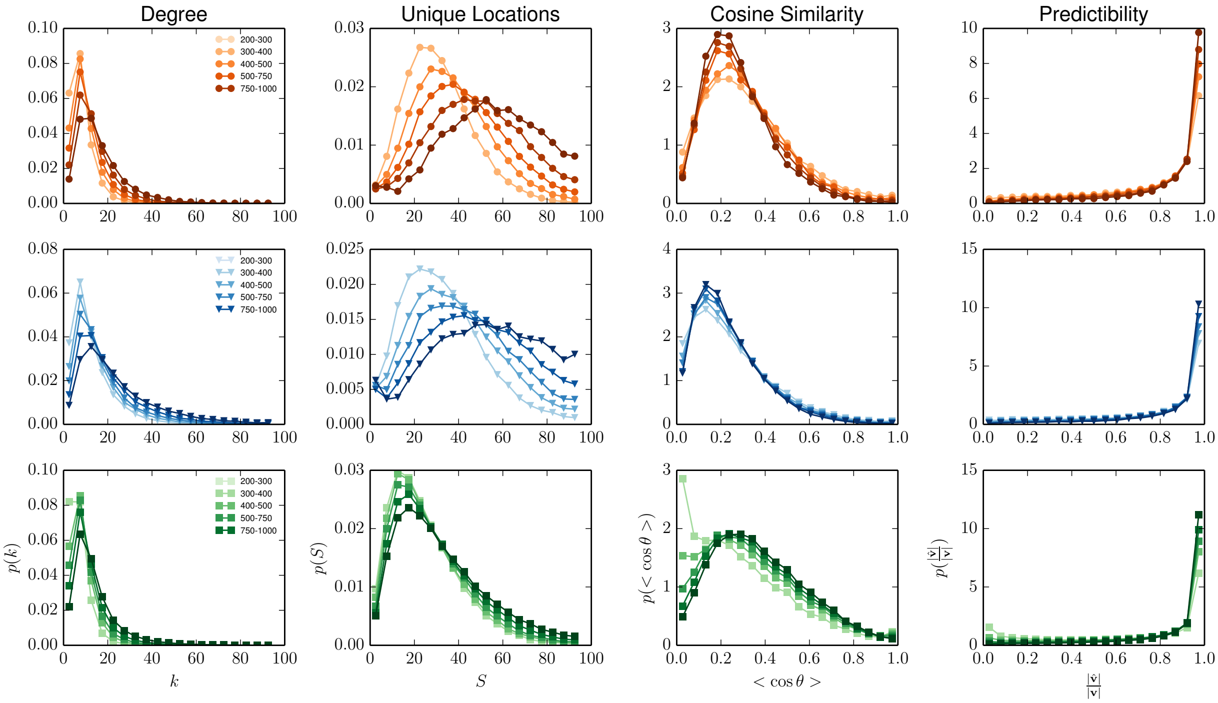

FIG. S1. Distributions of different variables (columns) for each of the three regions (rows) for groups of users with different numbers of total calls. To ensure that measurements are not simply artifacts of differences in the amount a users interacts with their phone, we plot distributions of variables for groups of users with different activity levels. Users are binned first according to the number of records they have in the data set, then distributions of various mobility and social metrics are plotted for each user group. In general calling frequency does not affect these distributions with the exception of the number of unique locations visited where the mean is shifted right for users with more calls.

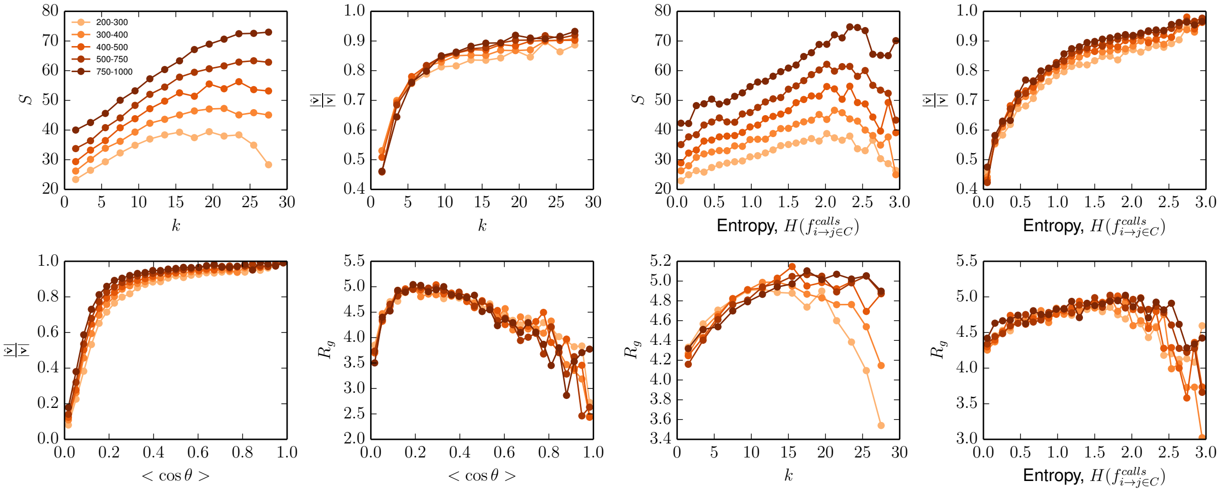

FIG. S2. Various correlations metrics related to social behavior and mobility while controlling for the number of calls made by each user. Again, we bin users based on the number of records they have in our data set and then measure correlations between social and mobility metrics. We find, as was the case with distributions, these correlations are unaffected by sampling frequency.

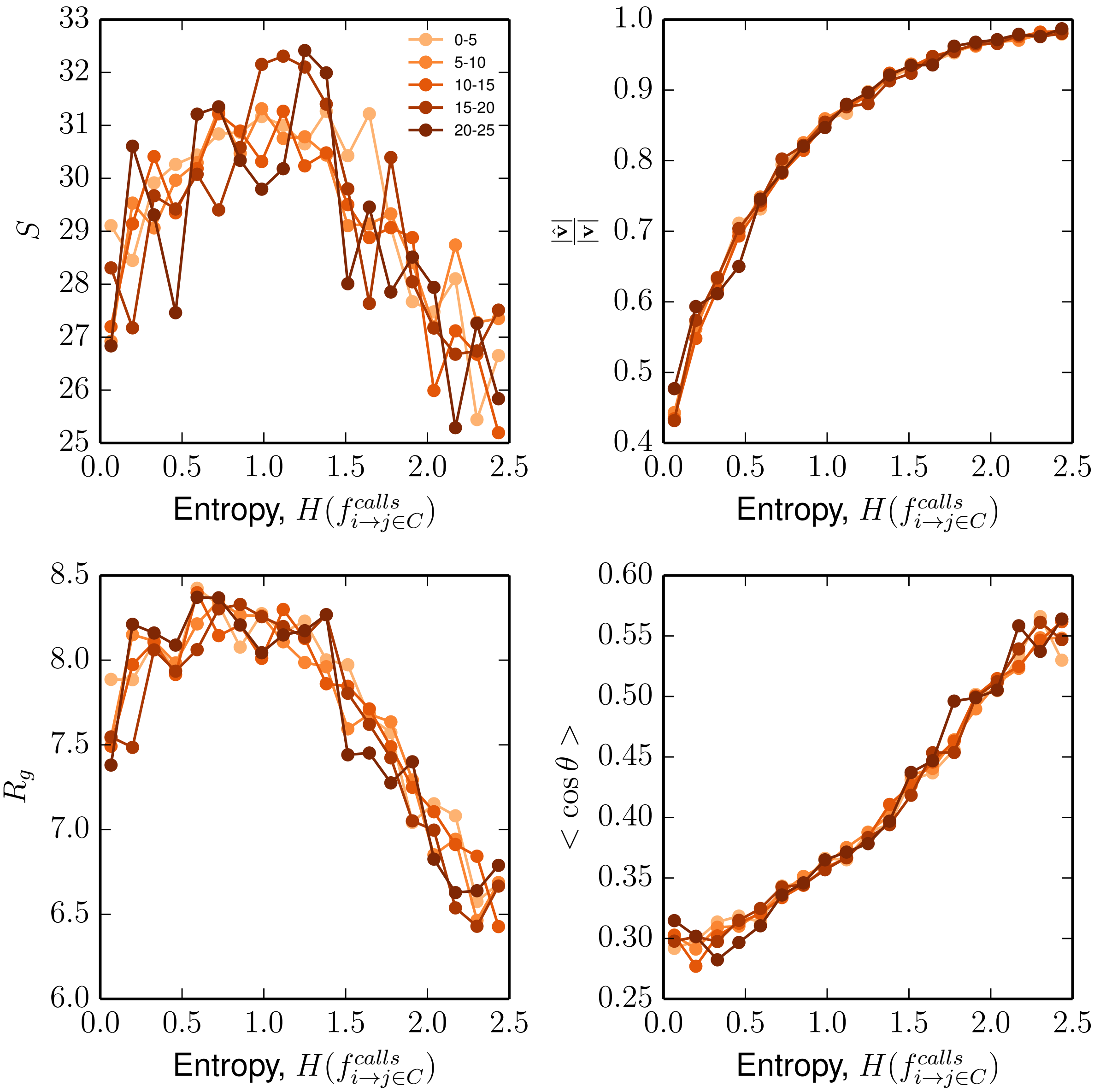

FIG. S3. Correlation between the entropy of a node's call frequency distribution to contacts and mobility variables may be affected by the degree of each node. Measures such as entropy and predictability will naturally be affected by the number of contacts each user has. For example if a user contacts for people, the maximum entropy of the distribution of call frequencies to those individuals will naturally be higher than a user who has few friends. To ensure our correlations are not artifacts of the number of contacts each user has, we plot these correlations for groups of users with the same degree and show that these relations still hold.

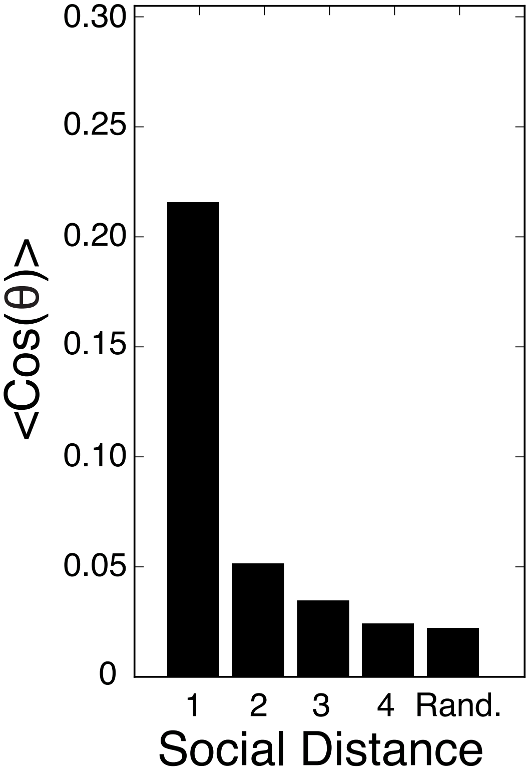

FIG. S4. Social distance and geographic similarity. Nodes who contact each other are far more similar to each other than two randomly selected nodes. Here we compute the average mobility similarity between nodes separated by a certain number of hops. Even for nodes separated more two or three hopes, we elevated levels of similarity when compared to two randomly selected nodes in the network.

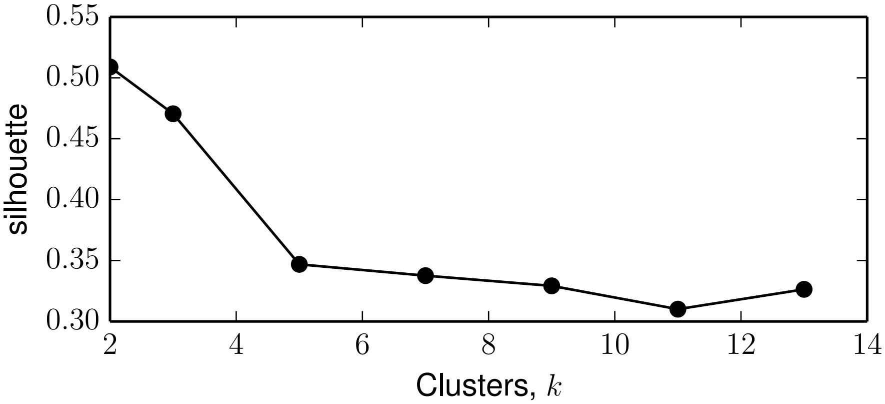

FIG. S5. The silhouette score for different numbers of clusters. The silhouette score is a measure of the ratio between intra- and inter-cluster variance that gives a rough measure of the quality of clustering results (higher is better). The score drops steadily from the chosen number of clusters, 3, indicating that little is gained by additional splitting.

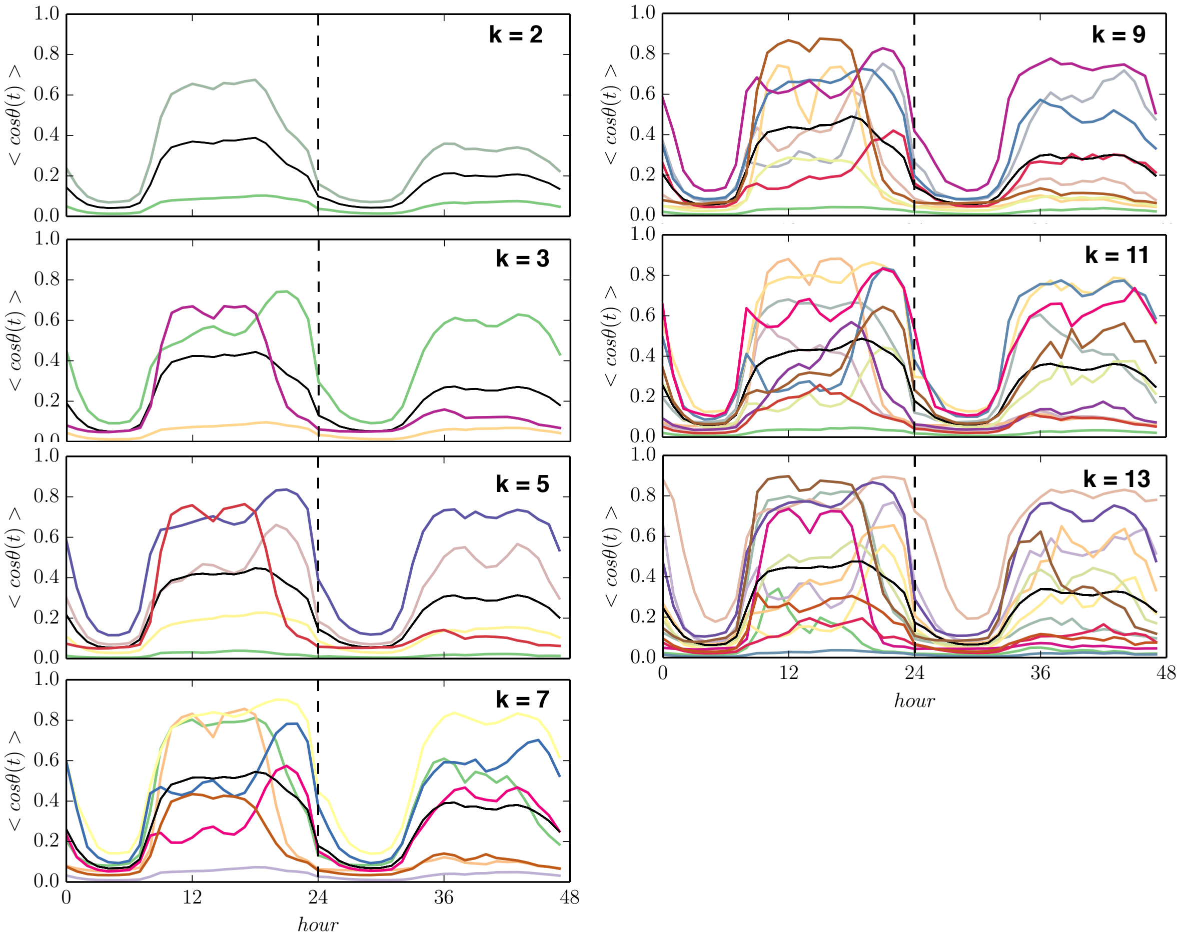

FIG. S6. k-Means clustering results for various values of k in city R1. We perform k-means clustering for multiple values of k as a manual check that our choice of 3 clusters is appropriate. In general, additional clusters appear to be variations of three main themes used in the main text.

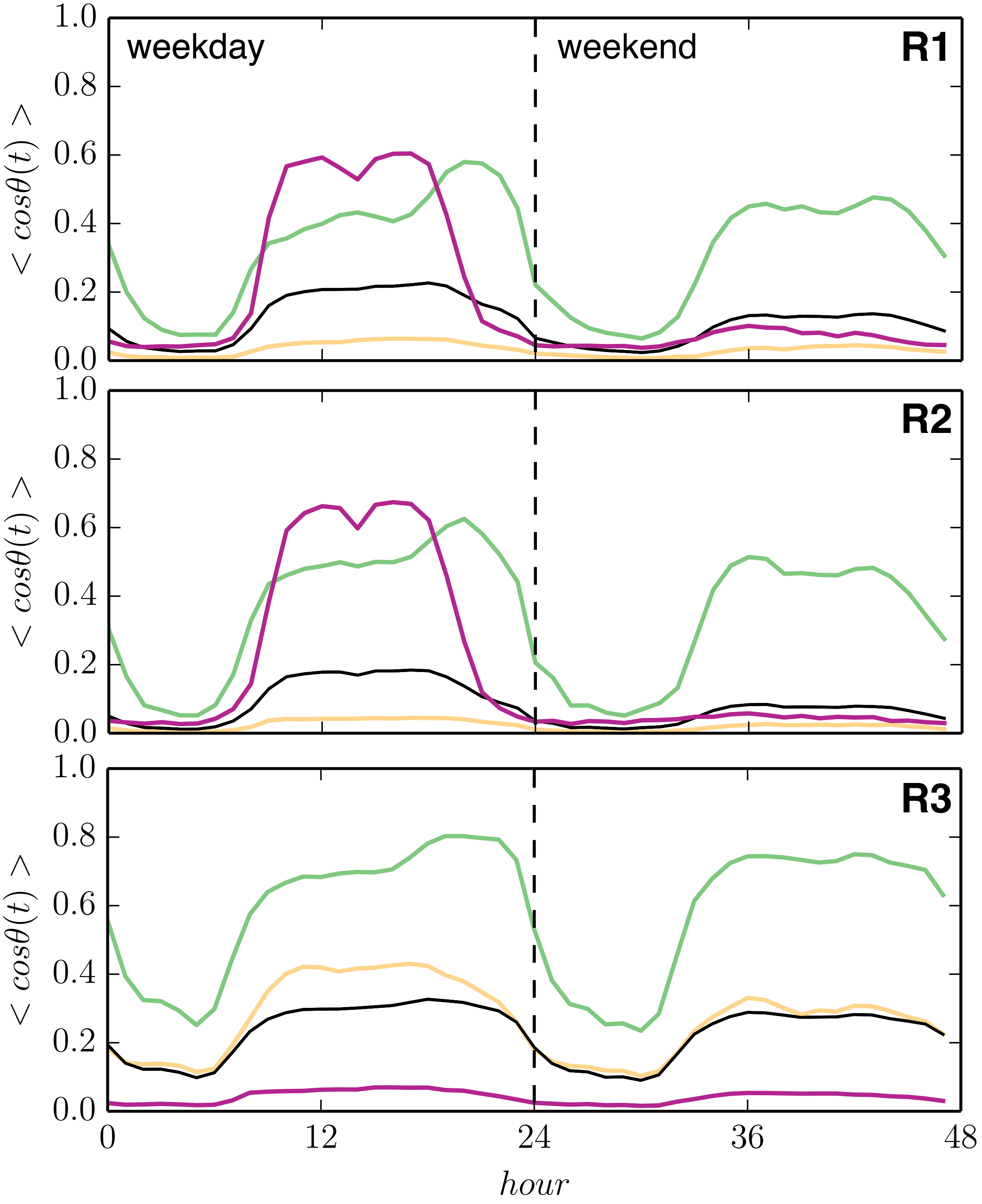

FIG. S7. Results from a hierarchical, agglomerative clustering algorithm with Ward linkage. This clustering method clusters nodes based on connecting data points together if they are within some distance of one another and then examining connected components. The clusters in each region closely match results from k-means, suggesting that our results are robust to the exact clustering algorithm used.

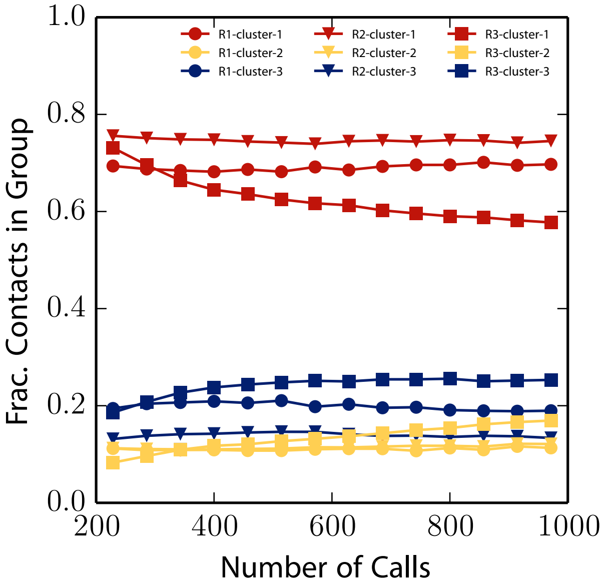

FIG. S8. The mix of a user's ego network versus the number of calls they make. To ensure that the relationship between the mix of a user's ego network and their mobility isn't simply due to the fact that user's with different numbers of records having different edge mixes, we plot the average makeup for suers with different numbers of calls. We find that regardless of a user's calling frequency, the makeup of their social contacts is stable.

MODEL COMPARISONS

GeoSim Model

To account for heterogeneity in how social different person's are, we allow each individual in our simulation to have a value of α drawn from a distribution. We run the GeoSim model with various distributions of α to determine which best approximate measurements from the data. Figure S9B shows the various distributions of α we simulate while Figures S9C and D show resulting similarity and predictability distributions when compared to empirical measurements. We find that exponential distributions $p(\alpha) \propto \exp(-\lambda\alpha)$ result in distributions that match the general trend of the empirical distributions. We expect the precise parameter, λ, to vary from city to city or culture to culture, but values between $0.1 and 0.3$ produce adequate results and are consistent with previous works that find roughly 15%-30% of trips are made for social purposes.

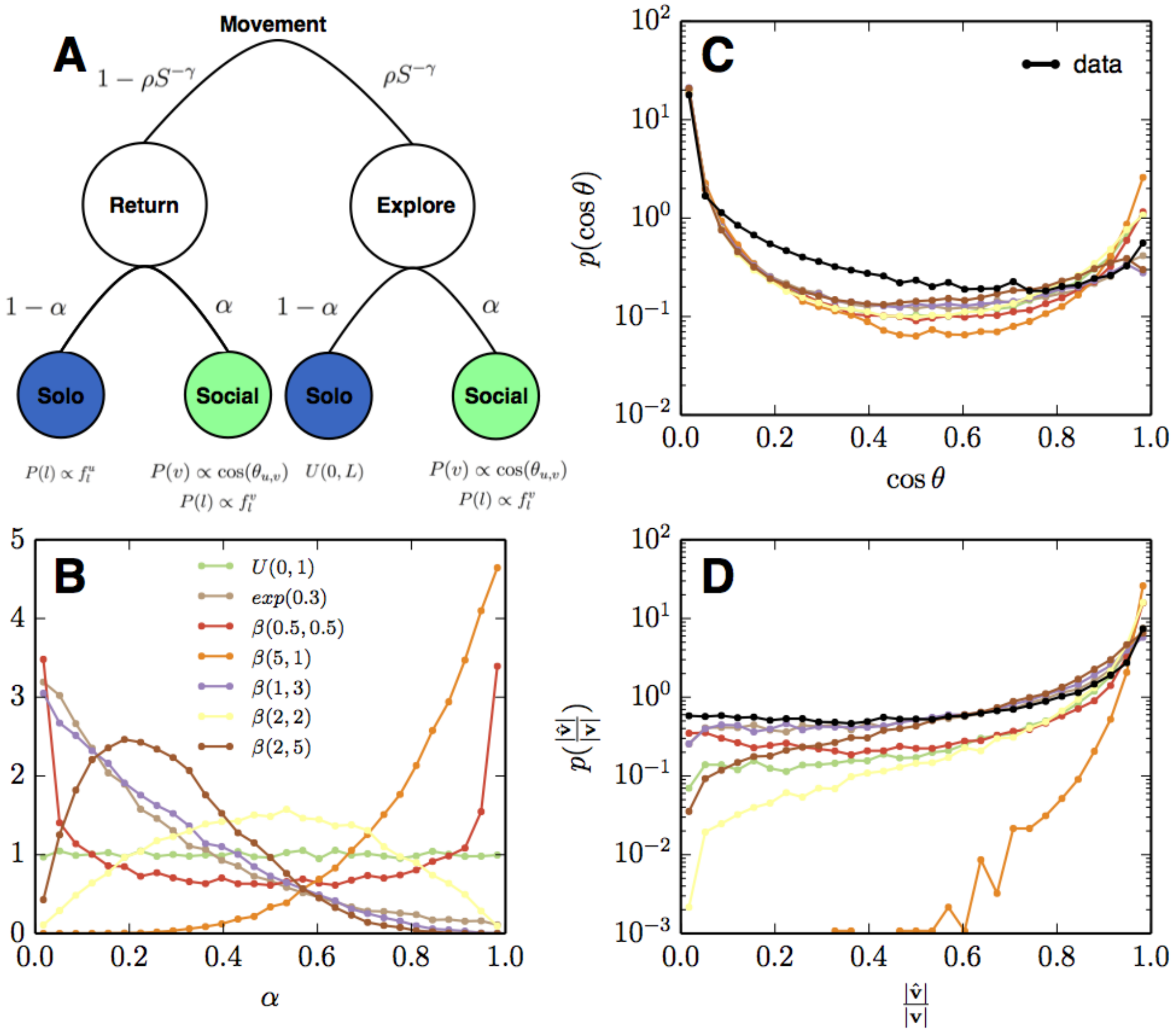

FIG. S9. Our extended mobility model. (A) A diagram showing the choices made by an individual in deciding where to move next. We compare simulation results for different values of social influence α to distributions of (B) similarity and (C) predictability found in real data. The bimodal similarity distribution is recovered for higher values of α while the predictability results suggest that this parameters may vary for individual to individual resulting in a mix among the whole population.

Individual Mobility Model (IM Model)

In their work, Song et al. [1] present a model for individual human mobility that we extend to the GeoSim model presented here. When $\alpha = 0$ for all individuals, we recover Song's model. The IM model relies on a 2 parameters, ρ and γ which control the propensity for a user to explore a new or return to a previously visited location. By varying these parameters, the authors can generate a range of mobility behaviors related to the rate of exploration and the frequency that users visit locations. We find that values of $\rho = 0.6$ and $\gamma = 0.6$ produce reasonable fits to both the exploration rates ($S(t)$) and the frequency that users return to locations f_k. We leave distributions such as the waiting time distribution measured by Song intact with $\beta = 0.8$.

Travel-Friendship Model (TF Model)

The model presented by Grabowicz et al. [2] proposes to model mobility and the growth of social networks simultaneously. In this model, geographic space is treated as a very small grid with cells δ on a side. At each time step, a user makes choices related to travel and friendship. For travel, an individual chooses to travel to the location of a random friend with probability p_v or with probability $1 - p_v$, chooses to jump to a new location. In the event of a jump, a distance is chosen based on the distribution measured by Song et al. and the individual then surveys all grid cells at that distance and chooses one to move to proportional to the population density of the cells. After a user has made a travel decision, they make choices related to friendship. For each other person within the individual's grid cell, a link between them is created with probability p, and with probability p_c a link is created with a random person anywhere. As opposed to other models, which assume the social network is fixed, the TF model builds the social network simultaneously with mobility choices. These dynamics reproduce social network distributions as well as distributions of distance between friends.

We implement the above model as described, but add one additional step to make it comparable to the CDR data used in this study and modeled by Song et al. The δ grid cells in the TF model are on the order of 100m × 100m in the original implementation. This is far smaller than the coverage areas of towers within a city save for the very dense downtown areas. To make the the data from the two models directly comparable, we simply assign a users location to the tower that covers the grid cell they have jumped to. This preserves all the original behavior of the TF model, while making it possible to perform a fair comparison.

Finally, the TF model has four parameters: δ, p, p_v, and p_c. We set these parameters to the middle of the ranges estimated by Grabowicz et al: $\delta = 0.001$, $p = 0.1$, $p_v = 0.15$, and $p_c = 10^{-3}$. We run this model for the same population size and length as the IM model and GeoSim models.

Source Code

A Python implementation of the GeoSim model has been made available at `http://humnetlab.mit.edu/wordpress/downloads/`.

* jltoole@mit.edu

[1] Song C, Koren T, Wang P, Barabási AL (2010) Modelling the scaling properties of human mobility. *Nature Physics* 6:818–823.

[2] Grabowicz PA, Ramasco JJ, Goncalves B, Eguiluz VM (2013) Entangling mobility and interactions in social media. p. 16.