Supporting Information

Deville et al. 10.1073/pnas.1525443113

Datasets

Mobile communication records, cataloged by mobile phone carriers for billing purposes, provide an extensive proxy of human movements and social interactions at a societal scale. Indeed, by keeping track of each phone call between two users and the spatiotemporal information about the users who initiated and received the call, mobile phone data offer information on both human mobility and social communication patterns at the same time, as we will detail hereunder.

In this project, we compiled a uniquely rich database consisting of three different datasets that are of a similar level of detail yet with different demographics, economic status, and scales:

Dataset $D1$ contains mobile phone calls between 1.3 million users over a period of 1 mo in 2006 from a European country (Portugal). For each phone call, the caller and the callee, both anonymized with a key (hash code); the time; the date; and the phone towers routing the communication are recorded. Only phone calls between users that called each other at least 5 times over a period of 18 mo are known. Furthermore, only the coordinates of the mobile phone towers are known; hence, the position of a user within the range of an antenna is unknown.

Dataset $D2$ covers a 6-mo period of mobile phone calls between 6 million anonymized users from a large European country. For each phone call, the caller, the callee, the time, and the towers routing the communication are recorded. Similarly to $D1$, only the coordinates of the mobile towers are known; hence, the position of a user within the range of an antenna is unknown.

Dataset $D3$ covers a period from 2005 to 2009 and is made of all transaction logs of all mobile phone activity that occurred in an African country (Rwanda) over the 5-y period. The data originate from the largest mobile phone operator in that country and contain about 1.5 million phone calls. The logs include the date, the time, and the mobile phone towers routing the call for each of the phone calls and are again anonymous. Again, only the coordinates of the mobile towers are known; hence, the position of a user within the range of an antenna is unknown.

Inferring Mobility and Social Fluxes

Mobility Fluxes. For each phone call, the position of the tower routing the call is known for the caller. Because we know the location of each tower, we know the location of the user was within the range of the tower's service area. By looking at each consecutive phone call made by a user, we can thus reconstruct the user's jumps between two consecutive locations where his calls were initiated. By aggregating all movements for all users, we can thus obtain the total number of jumps from any tower i to any tower j ($T^M_{i,j}$). All jumps made outside continental territories (i.e., islands) were not taken into account. The jumps do not exceed ~1,000 km, ~400 km, and ~100 km for datasets $D1$, $D2$, and $D3$, respectively, due to national frontiers and coverage limitations driven by geographical constraints in the country. We consider the number of jumps between two locations as the mobility fluxes between them.

Social Fluxes. For each phone call, the position of the tower routing the call is known for both the caller and the callee. By considering all phone calls, we thus know the total number of calls from a tower i to a tower j ($T^S_{i,j}$). We consider the number of phone calls between two locations as the social fluxes between them.

Jump Size Distribution at Fixed Interevent Time

It is known that the distribution of interevent times between two consecutive calls (locations) from the same user is heterogeneous (2). It is thus important to investigate if the observed displacement statistics (the jumps) are affected by this characteristic.

To simulate location traces left by a phone on a regular basis (instead of those due to calls) using data available to us, we calculate location displacement between a fixed time interval instead of two consecutive phone calls. More specifically, we use our dataset $D1$ and calculate the jump size distribution $P^M(r)$ for displacements separated by a time $\Delta T \pm 0.05\Delta T$. We systematically vary ΔT from 1 h to 1 d (Fig. S1). We find the distributions collapse for different choices of ΔT, suggesting that the use of consecutive calls serves as a good proxy for movements. Also, the curves can be well approximated by a power law consistent with our previous results. Note our results are bounded by the maximum distance a user can travel during ΔT, thus explaining the differences in the tail part of the distribution.

Distribution of Social and Mobility Fluxes for $D2$ and $D3$

In Figs. S2 and S3, we present results obtained for the datasets $D2$ and $D3$ regarding the distributions of the social fluxes $P^S_T(T|r')$ and $P^S_T(T|r)$ (Fig. 3 A and D for $D1$) and mobility fluxes $P^M_T(T|r')$ and $P^M_T(T|r)$ (Fig. 3 B and E for $D1$) for pairs of locations that are of similar distances (r and r'). We also show how flux distributions collapse for both datasets when they are rescaled with their average flux, $\langle T(r)\rangle$ or $\langle T(r')\rangle$ (Fig. 3 C and F). The same procedure for the dataset $D1$ is applied to $D2$ (Fig. S2) and $D3$ (Fig. S3). We again find that the fluxes for each group still follow a fat-tailed distribution, indicating there also exists much heterogeneity in fluxes among locations within similar distances for both $D2$ and $D3$. Locations that are nearby (small r' or r) tend to have higher fluxes, corresponding to higher intensity in both communications (Figs. S2 A and D and S3 A and D) and movements (Figs. S2 B and E and S3 B and E), corroborating our results for $D1$. Indeed, the curves shift to the right as r' (or r) decreases, indicating the probability for faraway location pairs to have large fluxes is much lower. Once we rescale the flux distributions with the average flux, $\langle T(r')\rangle$ or $\langle T(r)\rangle$, we find all of the curves collapse into a single curve, demonstrating again a single universal flux distribution characterizes both social communication and human movement fluxes, independent of distance (Figs. S2 C and F and S3 C and F).

Correlation Between Social and Mobility Fluxes with Geodesic Distance

As developed in the manuscript for the rank-based distance, we here analyze the correlations between the social fluxes $T^S_{i\to j}(r)$ and mobility fluxes $T^M_{i\to j}(r)$ in the case of the geodesic distance. We group location pairs (i and j) based on their distance and measure the relationship between $T^S_{i\to j}(r)$ and $T^M_{i\to j}(r)$ for each group ($r=1$, $r=5$, $r=10$, $r=50$, and $r=100$ in Fig. S4 A–E). In these scatterplots, each gray dot represents a pair of locations, and its x–y coordinates correspond to the mobility $[T^M_{i\to j}(r)]$ and social $[T^S_{i\to j}(r)]$ fluxes from i to j for dataset $D1$.

Same as for the rank-based measures, we find again strong correlations between these two quantities regardless of how far away these locations are separated. To quantify this correlation, we measure the average social fluxes given the mobility fluxes at a certain distance, $\overline{T^S}(T^M|r)$ (colored symbols in Fig. S4 A–E), which is formally defined as

$$\overline{T^S}\left(T^M|r\right) \equiv \frac{\sum_{i\to j} T^S_{i\to j}\delta\left(T - T^M_{i\to j}\right)\delta\left(r - r_{ij}\right)}{\sum_{i\to j}\delta\left(T - T^M_{i\to j}\right)\delta\left(r - r_{ij}\right)}, \quad \textbf{[S1]}$$

where $\delta(x)$ is the delta function [$\delta(x) = 1$ when $x = 0$, and $\delta(x) = 0$ otherwise]. We find that the average social fluxes $\overline{T^S}(T^M|r)$ have again a power law scaling relationship with T^M, following

$$\overline{T^S}\left(T^M|r\right) = A(r)T^M(r)^{\theta_r}, \quad \textbf{[S2]}$$

where the scaling exponent $\theta_r < 1$ for different r, indicating social fluxes again scale sublinearly with mobility fluxes. The prefactor in Eq. **S2**, $A(r)$, corresponds to the shift along the y axis through Fig. S4 *A*–*E*. We find, as distance increases, the average social fluxes increase given the same volume of mobility fluxes. Rescaling $\overline{T^S}$ by r^{δ_r}, we find all curves collapse into a straight line (Fig. S4*F*), indicating $A(r) \sim r^{\delta_r}$.We repeated the same measurement for *D*2 and *D*3. We found that although each dataset is again characterized by a different set of θ_r and δ_r, Eq. **S2** (same as Eq. **7** in the main manuscript) holds consistently well across different datasets (Fig. S4 *F*–*H*), demonstrating the robustness of our findings for both the distance r and r'.

Derivation of the Scaling Relationship Between Exponents

As stated in the main manuscript, we find that the average social fluxes $\overline{T^S}(T^M|r')$ follow a power law scaling relationship with T^M, i.e.,

$$\overline{T^S}\left(T^M|r\right) = A(r')T^M(r')^{\theta_{r'}}, \quad \textbf{[S3]}$$

where $\theta_{r'} < 1$, indicating social fluxes scale sublinearly with mobility fluxes, independent of distance. As described in *Correlation Between Social and Mobility Fluxes with Geodesic Distance*, a similar result is obtained for geodesic distance metric r (Eq. **S2**).

The data collapses observed in Fig. 3 *C* and *F*, i.e.,

$$P^{S,M}_T(T|r') = \left\langle T^{S,M}(r')\right\rangle^{-1}\mathcal{F}\left(T^{S,M}/\left\langle T^{S,M}(r')\right\rangle\right), \quad \textbf{[S4]}$$

together with Eq. **S3** allow us to derive a new scaling relationship between different critical exponents. Indeed, the average social fluxes at distance r', $\overline{T^S}(r')$, can be obtained by integrating $\overline{T^S}(T^M, r')$ over T^M:

$$\overline{T^S}(r') = \int P^M_T\left(T^M|r'\right)\overline{T^S}\left(T^M, r'\right)dT^M. \quad \textbf{[S5]}$$

Substituting Eqs. **S4** and **S3** into Eq. **S5**, we have

$$\overline{T^S}(r') = \int \mathcal{F}(x)\overline{T^S_M}\left(\overline{T^M}(r')x, r'\right)dx \sim \overline{T^M}(r')^{\theta_{r'}}r'^{\delta_{r'}}\int x^{\theta_{r'}}\mathcal{F}(x)dx, \quad \textbf{[S6]}$$

where $x \equiv T^M/\overline{T^M}$ as a change of variable. As $\overline{T^S}(r') \sim \sum_{i\to j} T^S_{i\to j}\delta(r' - r'_{ij}) = P^S(r') \sim r'^{-\beta_{r'}}$, and similarly $\overline{T^M}(r') \sim r'^{-\alpha_{r'}}$, we have

$$r'^{-\beta_{r'}} = r'^{-\alpha_{r'}\theta_{r'}}r'^{\delta_{r'}}\int x^{\theta_{r'}}\mathcal{F}(x)dx. \quad \textbf{[S7]}$$

The tail behavior of $\mathcal{F}(x)$ indicates the integral in Eq. **S7** converges. Hence, Eq. **S7** leads to a scaling relationship,

$$\beta_{r'} = \alpha_{r'}\theta_{r'} - \delta_{r'}, \quad \textbf{[S8]}$$

connecting the exponent that characterizes social communications ($\beta_{r'}$) with the exponent characterizing human movements ($\alpha_{r'}$). Similarly, for geodesic distance metric r, we obtain

$$\beta_r = \alpha_r\theta_r - \delta_r. \quad \textbf{[S9]}$$

The scaling analyses performed here have their roots in the canonical statistical physics literature, namely, the scaling identities in phase transitions and critical phenomenon. The power law scaling behavior in the vicinity of a continuous transition is captured by a set of critical exponents $(\alpha, \beta, \gamma, \delta, \sigma, \eta, \ldots)$, characterizing various fundamental quantities such as free energy, specific heat, magnetization, susceptibility, etc. In the beginning, these critical exponents were measured independently and found to vary slightly across different materials. Later, we witnessed a burst of results demonstrating that these critical exponents are not independent but are in fact connected through what we now call scaling identities. The famous examples include Rushbrooke's identity, Widom's identity, Josephson's identity, and Fisher's identity (58).

Determination of Gravity Law's Parameters

The gravity law assumes that the mobility fluxes between a locations i of origin and a location j of destination can be expressed as a function of the two populations at the two locations (m_i and m_j) and the geodesic distance between them ($r_{i,j}$) as

$$T^M_{GM,i,j} = C\frac{m_i^\mu m_j^\kappa}{f\left(r_{i,j}\right)}, \quad \textbf{[S10]}$$

where $f(r) = r^\gamma$ (6, 20, 45, 59). By taking the logarithm on both sides we obtain

$$log\left(T^M_{GM,i,j}\right) = log(C) + \mu log(m_i) + \kappa log\left(m_j\right) - \gamma log\left(r_{i,j}\right). \quad \textbf{[S11]}$$

Using the observed mobility fluxes (T^M), we can then estimate the parameters through a least square regression, giving us $[log(C), \mu, \kappa, \gamma] = [-3.42, 0.67, 0.68, 1.32]$.

Epidemic Spreading Simulations

To compare the accuracy and usefulness of our rescaling formula, we simulated an SIS process commonly used in modeling disease spreading (54, 55) by following the observed mobility fluxes T^M and the rescaled social fluxes $\tilde{T}^S$ but also the mobility fluxes T^M_{gm} approximated by the well-known gravity model (20, 45).

We consider the process where each location i (mobile tower) is characterized by a constant population size N_i, equal to the number of distinct users present in the vicinity of the mobile tower over the period covered by the dataset *D*1. The total population in our system is thus given by $\sum_{i=1}^m N_i$, and the system equilibrated as the population is constant. In each location, users are classified according to their infectious state: they can be either infectious (I) or susceptible to be infected (S). The standard generalization of this spatial SIS model is given by

$$S_i + I_i \overset{\mu}{\to} I_i \quad \textbf{[S12]}$$

$$I_i \overset{\nu}{\to} S_i \quad \textbf{[S13]}$$

$$S_i \overset{A_{i,j}}{\to} S_j \quad \textbf{[S14]}$$

$$I_i \overset{A_{i,j}}{\to} I_j, \quad \textbf{[S15]}$$

where reaction **S5** indicates that susceptible users can become infectious at a rate μ and reaction **S6** corresponds to infected users recovering from the disease at a rate ν. In addition to the standard SIS dynamics, susceptible as well as infected users can randomly move between one location i to another location j as

described in reactions **S7** and **S8**. The probability rate of these movements from location i to j is governed by the probability rate $A_{i,j}$ defined as

$$A_{i,j} = \frac{\left(T_{i,j}T_{j,i}\right)^{-2}}{N_i}. \quad \textbf{[S16]}$$

Because the system is equilibrated, the flux of users from i to j must balance that of j to i (detailed balance condition):

$$A_{i,j}N_i = A_{j,i}N_j, \quad \textbf{[S17]}$$

which is fulfilled by Eq. **S18**.

In this case, the spatial SIS model can be defined as a set of m coupled ordinary differential equations (ODEs) for the infected people in each location (22, 60):

$$\partial_t I_i = \mu \frac{I_i}{N_i}(N_i - I_i) - \nu I_i + \sum_{j \neq i}\left[A_{j,i}I_j - A_{i,j}I_i\right], \quad \textbf{[S18]}$$

enabling us to compute the evolution of infected users in each location over time by solving these.

Denoting with n_i the number of users at location i, with a_i the area of location i and $m_i(t)$, $\tilde{m}_i(t)$, and $m_i^{GM}(t)$ the number of infected users at time t in location i when using T^M, $\tilde{T}^S$, and T^M_{GM}, respectively, we measure $m_i(t)/a_i$, $\tilde{m}_i(t)/a_i$, and $m_i^{GM}(t)/a_i$, i.e., the density of infected users estimated in each location i for the three cases (Fig. 5 *A–C* for $t = 17$). We find a remarkable agreement between our simulation and the real spreading patterns. Moreover, close up on the city of Porto reveals a superior accuracy of our model comparing with predictions from gravity model. To quantify the differences between the two methods, we measure

$$\tilde{e}_i(t) = \frac{m_{i,t} - \tilde{m}_{i,t}}{m_{i,t}} \quad \textbf{[S19]}$$

and

$$e_i^{GM}(t) = \frac{m_{i,t} - m_{i,t}^{GM}}{m_{i,t}}, \quad \textbf{[S20]}$$

corresponding to the relative error of infection rate in each location i at time t for both methods (Fig. 5 *D* and *E* at $t = 17$). The drastic difference between Fig. 5 *D* and *E* highlights the fact that lower $\tilde{e}_i(t)$ are observed comparing with $e_i^{GM}(t)$ in this particular example, again documenting the superior predictive power of our model.

To systematically assess and compare the accuracy of our results, we simulated 500 independent spreading processes following the same procedure described above but choosing randomly μ and ν parameters as well as the initial infected location and the number of infected users. For each simulation, we compute the mean values $\bar{\tilde{e}}(t)$ and $\overline{e^{GM}}(t)$ from Eqs. **S19** and **S20**, respectively, at different stages (time steps). We find that $\bar{\tilde{e}}(t)$ obtained from the 500 simulations are systematically lower than $\overline{e^{GM}}(t)$ across all stages of the spreading processes (Fig. 5*F*), demonstrating the practical relevance of our scaling relationship, effectively predicting mobility patterns using social communication records.

Normalizing the Time Steps of the Spreading Processes

In this section, we describe the procedure we use to compare spatial spreading processes whose initial conditions differs.

As formulated in Eq. **S18**, each spatial process is characterized by a set of m coupled ODEs. Each ODE corresponds to a spreading subprocess within a location, and each one of them reaches the steady-state after a different number of time steps (61, 62). Here we consider the global process to be at equilibrium when no more changes are observed for any of its subprocesses, i.e., $\max\ \partial_t I_i/N_i < 10^{-5}$.

As described in the main manuscript, we simulated 500 spatial spreading processes, each with parameters μ, ν, initial infected location, and initial number of infected users chosen randomly. Each process i will thus reach the equilibrium at a different time t_i^c. To compare their accuracy at different stages of the process, we normalize the time steps t_i of each process i by its time before equilibrium, i.e., t_i/t_i^c. As a result, a time step of $t_i/t_i^c = 0.5$ for any process i would correspond to half the time it takes to reach the equilibrium. This normalization is used in Fig. 5*F* to compare processes at similar stages.

Potential Limitations of Mobile Phone Datasets

For studies on mobility and social interactions, the mobile phone dataset is the most relevant dataset that is currently in existence. Indeed, at present, the most detailed information on human mobility across a large segment of the population is collected by mobile phone carriers. Mobile carriers record the closest mobile tower each time the user uses his or her phone. Other possible data sources include dollar bills, GPS, or check-in datasets from location-based social networking services, all of which suffer from well-known limitations that are resolved by mobile phone datasets. Indeed, dollar bills are carried by various individuals; hence, mobility inferred from them captures population-level aggregated movements instead of individual mobility. GPS tracks individual positions on a continuous basis with high precision, but it operates on a much smaller scale (typically hundreds of people) in contrast to millions of individuals' mobile phone data records. Check-in datasets only record mobility information when users report their positions voluntarily on subset of population who use the service, in contrast to mobile phones that objectively collect mobility information across a societal-scale population. For this reason, research on human mobility has literally exploded following the availability of mobile phone datasets, resulting in a number of rather fundamental papers. Furthermore, mobile phone datasets offer comprehensive information on phone calls and text messages, providing social network information in addition to mobility trajectories of each individual. Therefore, mobile phone datasets are excellent data sources to study simultaneously human mobility and social networks.

However, mobile phone datasets have a number of well-known limitations. Most notably, there are three aspects:

First, as mobile phones approximate a user's location by the tower that routed the call, the spatial resolution of the dataset is limited by the area covered by each tower, which typically ranges between 1 and 3 km. This is a spatial limitation of the data. Luckily, earlier research has extensively focused on this issue and documented that, at least for results we discussed, the results are not affected by this limitation.

Second, a user's position is only recorded when he or she makes a call or sends a text. However, human communications follow bursty patterns. This is the temporal limitation of the data. However, there are ample reasons to believe that our results are not affected by it. Mobility studies that compare mobility patterns obtained through mobile phone data and other continuous tracing technologies consistently find that the two are largely indistinguishable (2). These include GPS traces (2) as well as high-resolution mobile phone records (4, 5). Although we do not have direct access to these datasets, using our own datasets, we further calculated location displacement between a fixed time interval instead of two consecutive phone calls, in doing so artificially creating mobility traces that occur on a continuous basis. We find that results are remarkably stable as we vary the time interval systematically from 1 h to 1 d (*Jump Size Distribution at Fixed Interevent Time*). All these results suggest that although

mobility information is obtained from calling patterns in mobile phone datasets, call detailed record (CDR) data offer representative patterns of mobility, offering convincing reassurance that our results are not affected by this limitation.

Third, social network information is inferred based on calling patterns, yet calls using mobile phones can be ambiguous and hence may not represent true social relationships. This is the third limitation of the data. However, results by Eagle et al. (56) compared self-report survey data on mobility and social interactions with observational data obtained using mobile phones, demonstrating a high degree of accuracy (95%) in inferring friendship structures based on observational data alone.

Taken together, mobile phone datasets are the best and largest datasets for the type of study we conducted here. Although they have well-known limitations, extensive studies and results have demonstrated that our study is not affected by these limitations.

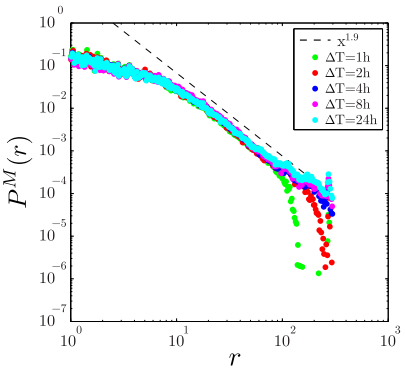

Fig. S1. Jump size distribution $P^M(r)$ for interevent time $\Delta T = 1,\ 2,\ 4,\ 8,$ and 24 h for the *D*1 dataset. A power law with exponent $r^{-1.9}$ provides a guide to the eye. We observe that the curves are bounded by the maximum distance a user can travel during their corresponding interevent time for $\Delta T < 4$ h or the maximum distance enforce by geographical constraints in that country.

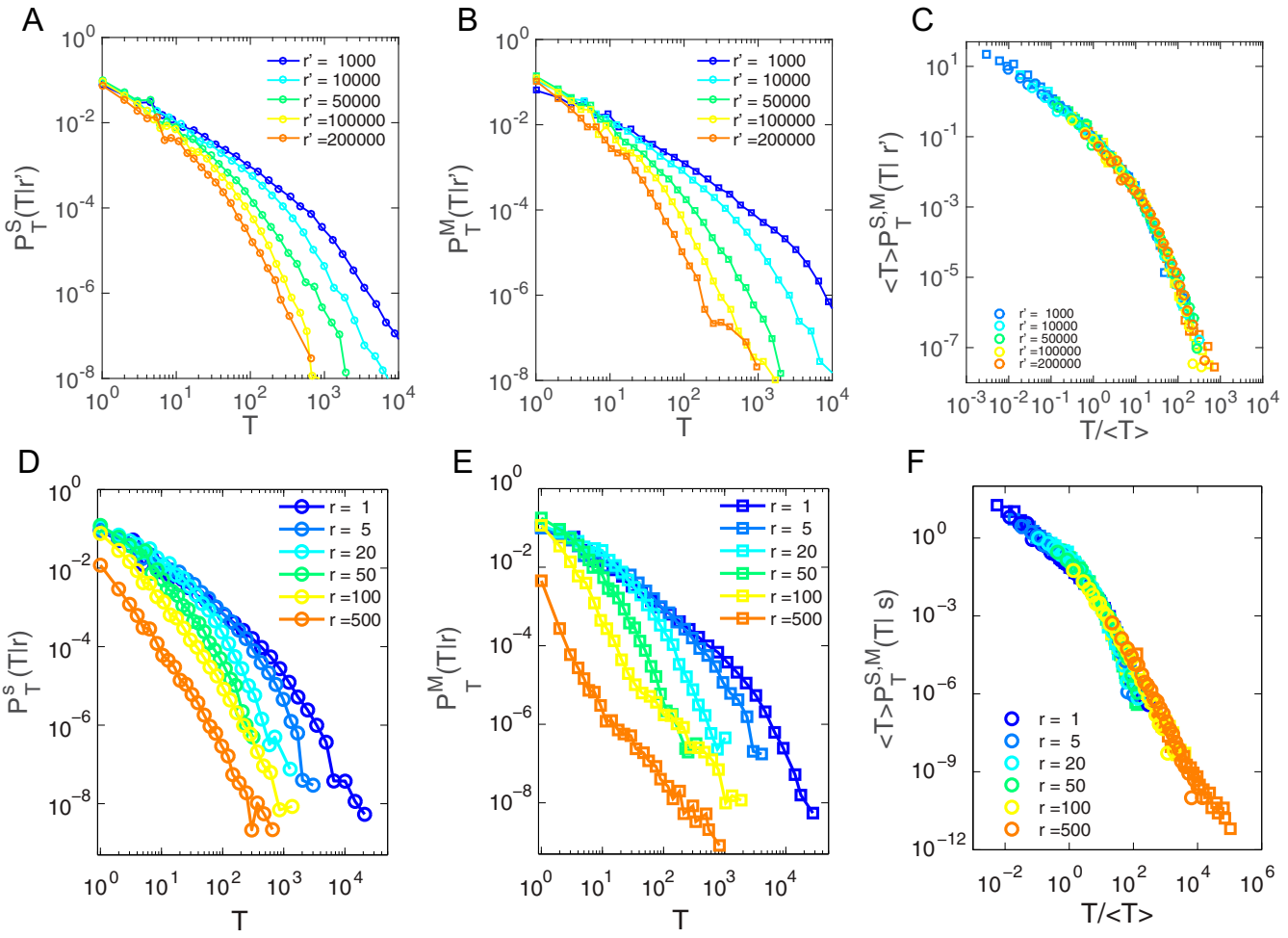

Fig. S2. (*A*) Communication fluxes $P^S(T|r')$ conditional to the rank r' for dataset *D2*. (*B*) Mobility fluxes $P^M(T|r')$ conditional to the rank r' for dataset *D2*. (*C*) Mobility and communication fluxes, denoted by circles and squares, respectively, collapse after rescaled by $\langle T\rangle$ for dataset *D2* for the rank-based distance. (*D*) Communication fluxes $P^S(T|r')$ conditional to the distance r for dataset *D2*. (*E*) Mobility fluxes $P^M(T|r)$ conditional to the distance r for dataset *D2*. (*F*) Mobility and communication fluxes, denoted by circles and squares, respectively, collapse after rescaled by $\langle T\rangle$ for dataset *D2* for the geodesic distance.

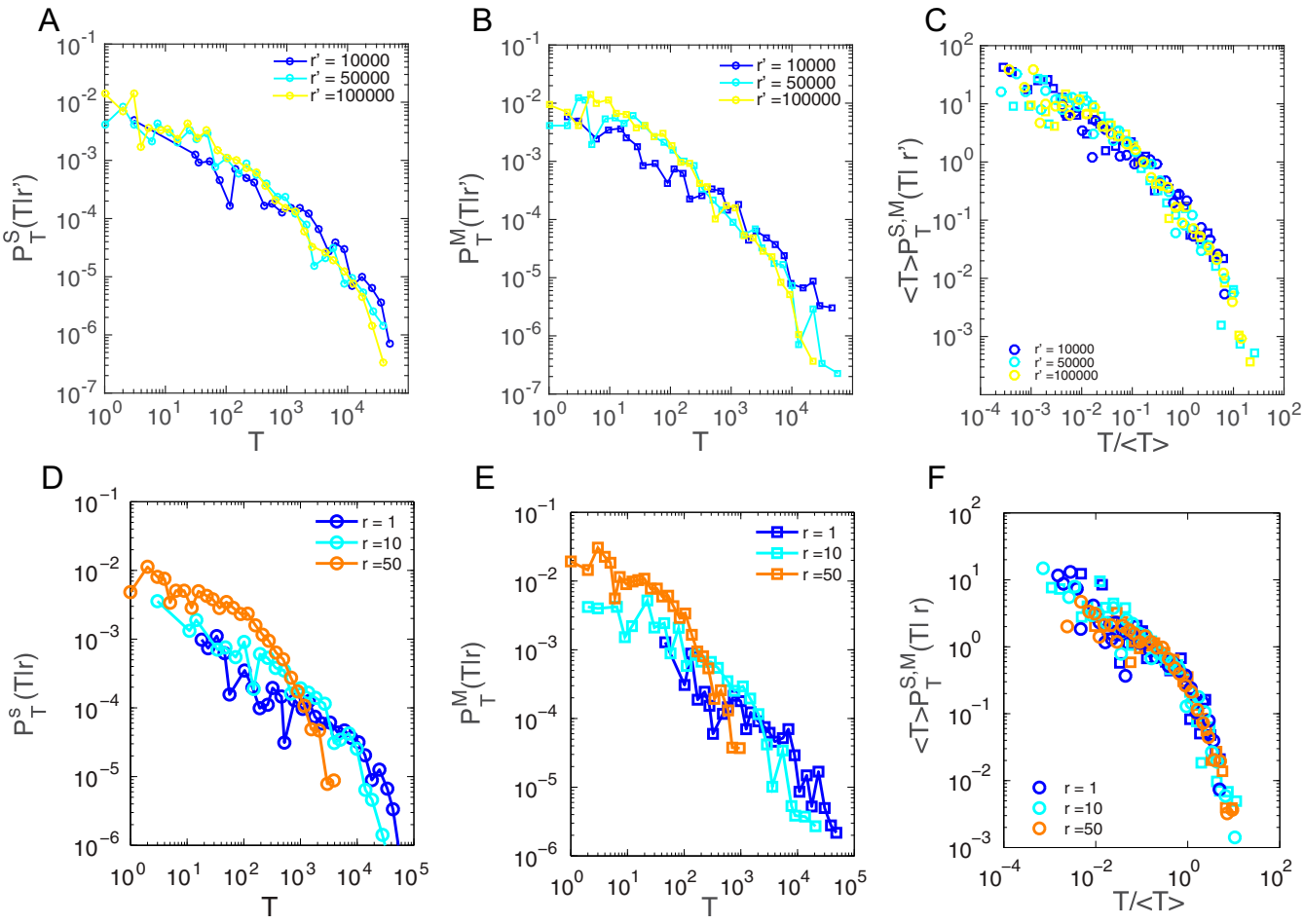

Fig. S3. (*A*) Communication fluxes $P^S(T|r')$ conditional to the rank r' for dataset *D*3. (*B*) Mobility fluxes $P^M(T|r')$ conditional to the rank r' for dataset *D*3. (*C*) Mobility and communication fluxes, denoted by circles and squares, respectively, collapse after rescaled by $\langle T \rangle$ for dataset *D*3 for the rank-based distance. (*D*) Communication fluxes $P^S(T|r)$ conditional to the distance r for dataset *D*3. (*E*) Mobility fluxes $P^M(T|r)$ conditional to the distance r for dataset *D*3. (*F*) Mobility and communication fluxes, denoted by circles and squares, respectively, collapse after rescaled by $\langle T \rangle$ for dataset *D*3 for the geodesic distance.

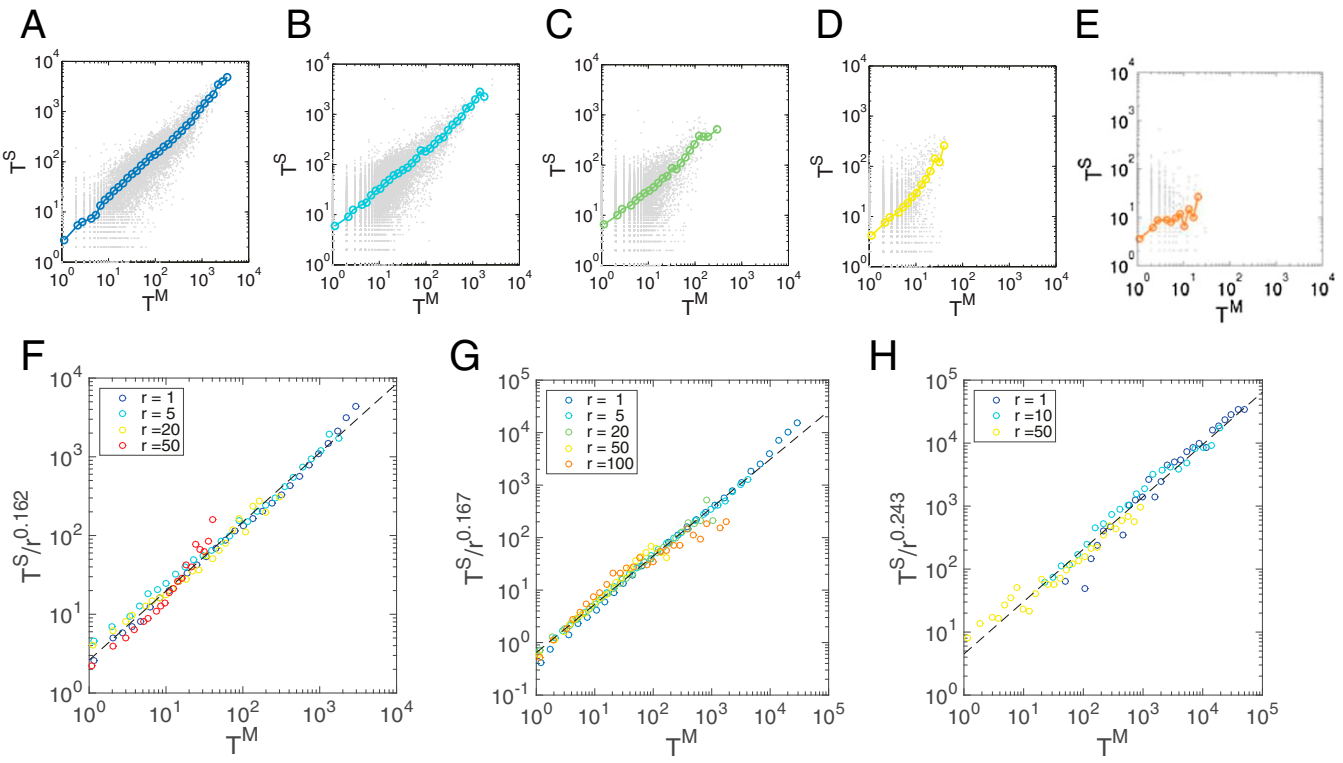

Fig. S4. Correlation between $T^S_{i\to j}(r)$ and $T^M_{i\to j}(r)$ for pair of locations (gray dots) separated by a distance of (*A*) $r = 1$ km, (*B*) $r = 5$ km, (*C*) $r = 20$ km, (*D*) $r = 50$ km, and (*E*) $r = 100$ km for *D*1. We find all curves collapse into a straight line when $\overline{T^S}$ is rescaled by r^{δ_r} for (*F*) *D*1, (*G*) *D*2, and (*H*) *D*3.