SCIENTIFIC REPORTS
natureresearch

OPEN

Networks and long-range mobility in cities: A study of more than one billion taxi trips in New York City

A. P. Riascos[1,3] & José L. Mateos[1,2,3]

We analyze the massive data set of more than one billion taxi trips in New York City, from January 2009 to December 2015. With these records of seven years, we generate an origin-destination matrix that has information of a vast number of trips. The mobility and flow of taxis can be described as a directed weighted network that connects different zones of high demand for taxis. This network has in and out degrees that follow a stretched exponential and a power law with an exponential cutoff distributions, respectively. Using the origin-destination matrix, we obtain a rank, called "*OD rank*", analogous to the page rank of Google, that gives the more relevant places in New York City in terms of taxi trips. We introduced a model that captures the local and global dynamics that agrees with the data. Considering the taxi trips as a proxy of human mobility in cities, it might be possible that the long-range mobility found for New York City would be a general feature in other large cities around the world.

The study and understanding of human mobility in cities is an important and challenging problem since more than half of the world population lives in urban areas[1]. Nowadays human mobility can be explored in detail thanks to the digital traces people leave on mobile/digital platforms[2,3]. Identifying global emerging patterns for human mobility is important in topics like urban planning, transport systems, the influence of the spatial distribution of a city in the mobility[4–7], and the encounter or contact networks that emerge[8]. In addition to all these aspects lying in the field of complexity and cities, we have the science of networks with well-established tools and methods to describe complex systems[9–11]. In many cases, networks provide an important framework to study transportation modes and their interactions[12,13].

Several studies have revealed that human mobility follows a long-range dynamics, akin to Lévy walks, as has been shown before as a common strategy in many animal species and humans[3,14]. In the context of networks, Lévy flights were introduced in[15] revealing that long-range displacements increase the capacity to reach efficiently to any site of the network by inducing the small-world property through the dynamics. This process has been explored in different cases as diverse as fractional diffusive transport[16–19], the dynamics on multiplex networks[20], human mobility[8,21], semi-supervised learning[22], among others[19,23–27].

In this research, we analyze the spatial activity of taxis as a proxy for human mobility in urban areas. From publicly available datasets, we generate an origin-destination (OD) matrix for trips during a period of seven years from January 2009 to December 2015. We identify zones with a high demand of this service and in this way, the movement of taxis can be described as a directed weighted spatial network with nodes representing high demand zones and links defined by the number of trips between two zones. In addition, we have geographic coordinates for all the nodes and the respective distances between them; as a result, the system can be described as a spatial network[28]. With all this information, available through the analysis of trip records, we study the spatial activity of taxis as a dynamical process in this particular structure. Several authors have explored spatio-temporal patterns in the mobility of taxis in different urban areas[29–31]. The system of taxis in New York City has been studied with different methods; in particular, considering the complete routes followed by the taxis on the street network[32–35].

To clarify the connection between mobility and networks, let us illustrate some ideas in connection with the relation between directed weighted networks and human mobility. In Fig. 1 we depict a schematic illustration of agents moving between $N = 10$ specific regions denoted as squares in a two-dimensional plane. In this reduced example, we have $\mathscr{T}$ =1000 trips and the values T_{ij} (for $i, j = 1, 2, \ldots N$) denote the number of trips between two

[1]Instituto de Física, Universidad Nacional Autónoma de México, Apartado Postal 20-364, 01000, Ciudad de México, Mexico. [2]Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Apartado Postal 04510, Ciudad de México, Mexico. [3]These authors contributed equally: A. P. Riascos and José L. Mateos. email: aperezr@fisica.unam.mx; mateos@fisica.unam.mx

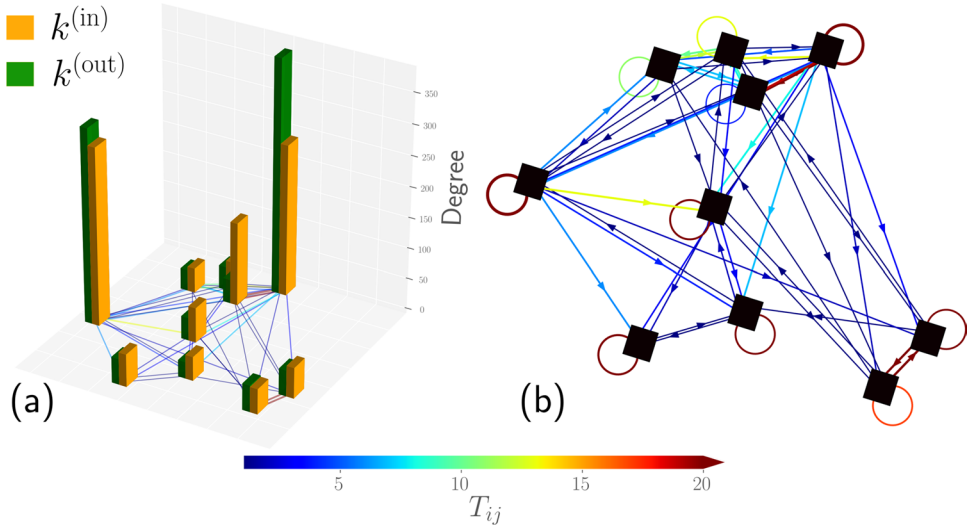

Figure 1. Schematic illustration of mobility as a spatially embedded directed weighted network. We show $N = 10$ square zones in the plane representing particular regions where agents can start or end a trip; we simulate $\mathscr{T} = 1\ 000$ trips of agents between these locations. (**a**) Bar representation of the total number arrivals $k^{(\mathrm{in})}$ and the number of departures $k^{(\mathrm{out})}$. (**b**) Diagram of the system expressed as a directed network, we represent with colors the number of trips T_{ij} between sites i and j. In the study of human mobility, this information is expressed as an origin-destination matrix $N \times N$ with elements T_{ij}. In particular, the directions of the links are depicted by arrows and self-loops represent the number of trips with the same origin and destination.

regions. In Fig. 1(a) we represent with bars the values obtained for the number of trips that arrive or depart each zone; in addition, colored lines denote the number of trips T_{ij}. In Fig. 1(b) we represent the complete structure described by the origin-destination matrix as a directed weighted network: the links have directions represented by an arrow and, with different colors in the lines, we depict the respective number of trips. Furthermore, we show with self-loops (i.e., a line connecting a zone with itself) the number of trips that start and end in the same zone, determined by the diagonal elements in the origin-destination matrix. In addition, in the study of mobility, the resulting structure is a spatial network and all the positions of the nodes are important, for instance, to determine the distance between two zones. This example shows the vast amount of information that is captured in the origin-destination matrix and its direct relation to a network, allowing us to use the full potential of network science to study mobility.

The paper is organized as follows. In the first part, we identify high demand zones and generate origin-destination matrices describing the global activity of taxi's flow. Then, we calculate the transition probabilities between high demand zones. We introduce a rank, called "*OD rank*", analogous to the page rank of Google. We also implement a model that describes the spatial activity of taxis and verify the predictions of this model with the real data through Monte Carlo simulations. Our findings reveal a well defined mathematical structure for the spatial mobility in urban areas with a dynamics that combines local displacements with a particular type of long-range movements. The methods introduced are general and can be used as a framework for the study of different transportation systems in cities.

Results

Activity between zones with high demand. We explore taxi trip records taking into account the administrative boundaries including the five boroughs of New York City[36]. As a result, for the seven years studied, we have $\mathscr{T} = 1\ 148\ 052\ 357$ taxi trips (see the Methods section for a detailed description of the datasets explored). In the following, we study this volume of data by using a grid with 500×500 square zones with dimensions 100 m $\times$ 100 m. Once this grid is defined, we examine the zones contained in the administrative boundaries of New York City. In Fig. 2, we present a map generated with the information of origin and destinations reported in the datasets. For each square zone defined before, we count the number of trips according to the registers of longitude and latitude of the initial and final locations of each trip for taxi registers from 2009 to 2015. The results depicted in Fig. 2 give us a first insight into the global activity of taxis. We can identify a high demand of this service in Manhattan, also the high activity in the John F. Kennedy (JFK) International Airport and how by exploring the origins of the trips we can detect some features of the street network in New York City. On the other hand, we can see in Fig. 2(b) that the destinations are less localized in specific zones observing that in the Bronx, Brooklyn and Queens boroughs the number of taxis arrivals is more uniform in comparison with the origins in Fig. 2(a). This particular feature reveals how taxi transportation manages to permeate almost all the regions of the city.

In Fig. 2, we can identify zones in New York City with low demand for taxis or where only a reduced number of taxis arrives. Even considering the counts in seven years of activity, we can identify zones with dimensions

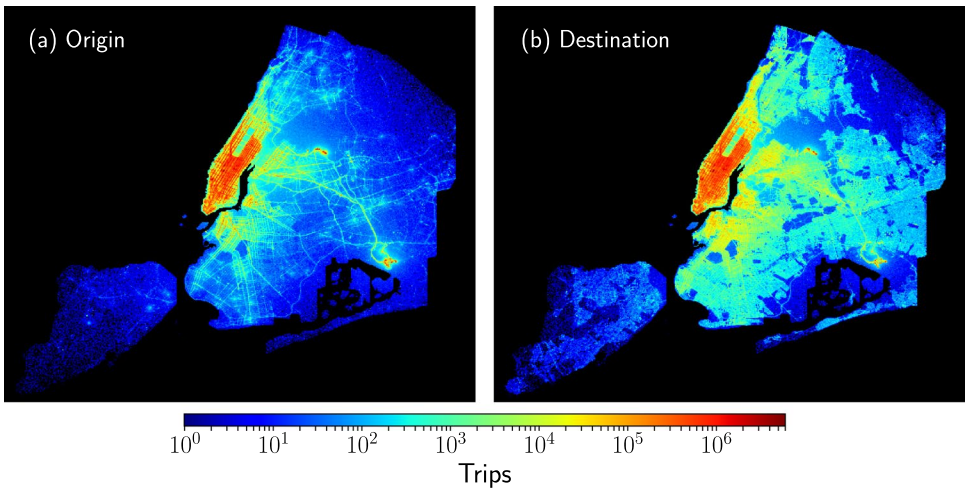

Figure 2. Origins and destinations of taxi trips in New York City. In this analysis, we divide the city in 100 m × 100 m square regions and, for each region, we count the number of taxi trips considering the registers of longitude and latitude of the initial and final locations of each trip. The results are presented in (**a**) for the origins and (**b**) for the destinations of taxis. The colorbar indicates the number of trips in each zone; regions outside the boundaries of New York City are presented in black. We analyze $\mathscr{T} = 1\ 148\ 050\ 837$ trips from taxi trip records between January 2009 to December 2015. In this representation of the data, using only the information of origins and destinations of taxis, we can see in detail the spatial complexity of New York City and how the street network emerges from the large number of trips analyzed.

Year	Fraction original database (%)	N	Displacements $\mathscr{T}$	$0 \leq d \leq 1.8$ Km (%)	$d > 1.8$ Km (%)
2009	91.76	4 456	153 389 115	44.69	55.31
2010	91.71	4 465	150 327 196	43.87	56.13
2011	91.70	4 558	156 962 079	42.92	57.08
2012	91.70	4 642	158 714 293	42.44	57.56
2013	91.65	4 645	154 833 137	42.69	57.31
2014	91.08	4 612	146 484 526	43.39	56.61
2015	90.22	4 353	128 984 657	43.21	56.79

Table 1. Analysis of the spatial activity of taxi trips in New York City considering zones with a high demand for this service. By using the rule that at least $\mathscr{M} = 1\ 000$ trips depart and arrive from a zone in a year, we obtain the number N of high demand zones. In addition, we present the total number of trips $\mathscr{T}$ between zones and the fraction of the original dataset that each number of trips represents. For the trips analyzed, we show the fraction of local trips with geographical distances d in the interval $0 \leq d \leq 1.8$ Km and the fraction of long-range movements with $d > 1.8$ Km.

100 m × 100 m for which less than 10^3 taxi trips arrive or depart. This is a small number in comparison with the values of zones with a high demand for which we observe more than a million arrivals or departures. Much of these zones are located in Manhattan but also other zones of the city. In what follows, we study the flow of taxis between zones with high demand and we will describe the global spatial dynamics as a directed weighted spatial network. All the zones in our study are defined by a square with dimensions 100 m × 100 m and, for each year, we classify a region as a high activity zone if, in this specific part of the city, the number of arrivals and departures are at least 10^3. In this way, the minimum number of arrivals at a high activity zone is at least 10^3 trips, and the same rule applies to the number of taxis leaving this region. This limit is reasonable due to the high quantity of trip records explored per year in the complete database. In addition, by using this rule we reduce possible errors produced by the inaccuracy in the origin and destination coordinates. By applying the criteria described before to the taxi trips in 2015, the number of high demand zones for this year is 4 353 and the total number of trips between these zones is $\mathscr{T} = 128\ 984\ 657$ that represents a 90.22% of the original database described in the Methods section. We found similar values for the trips from 2009 to 2014. The results for the number of high demand zones N and the total number of trips $\mathscr{T}$ are presented in Table 1.

Now, we define origin-destination matrices describing the flow of taxis between high-demand zones. In this way, the global dynamics can be explored and treated as a directed weighted network; in particular, a spatial network for which the nodes represent zones of high demand and the links with weights can represent several quantities like the flow of vehicles, the geographical distance between nodes, among other values[28].

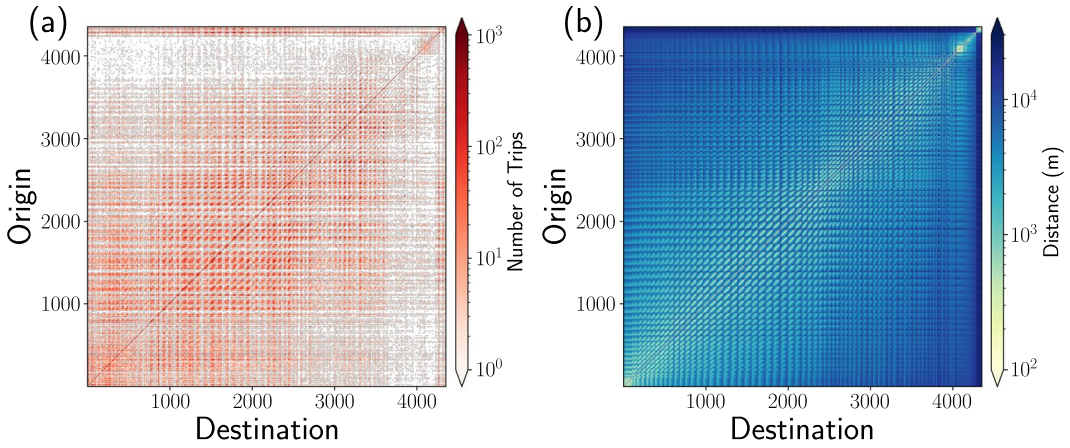

Figure 3. Global activity of taxis between zones of high demand for this service in New York City. We analyze the movement of taxis trips made in 2015 and, from the study of a grid with square zones with 100 m × 100 m, similar to the one presented in Fig. 2, we identify zones of high demand of taxis considering that at least 1 000 trips have departed or arrived from a zone. We found with this criterion $N = 4\,353$ high demand zones. In (**a**) we present the origin-destination matrix for taxi trips moving between zones, the respective colorbar codifies the trip counts. In (**b**) we present the geographical distance between origin and destination zones; the values of the distance are represented in the colorbar.

For each year, we calculate an origin-destination matrix for which the elements T_{ij} represent the number of taxi-trips from zone i to zone j, where $i, j = 1, 2, \ldots, N$ denote the square zones of high demand with dimensions 100 m × 100 m. In addition to the elements of the origin-destination matrix, it is important the in-degree defined as

$$k_i^{(\mathrm{in})} = \sum_{\ell=1}^{N} T_{\ell i} \tag{1}$$

that gives the total number of vehicles arriving at the zone i. We also have the out-degree determined by the relation

$$k_i^{(\mathrm{out})} = \sum_{\ell=1}^{N} T_{i\ell}, \tag{2}$$

that counts the total number of trips originated from zone i. On the other hand, to explore the spatial activity is important to have information about the geographical distances between zones. This information is included in a $N \times N$ distance matrix $\mathbf{D}$ with elements d_{ij} with the geographical distance between i and j. In addition, the degrees in Eqs. 1–2 satisfy:

$$\sum_{i=1}^{N} k_i^{(\mathrm{out})} = \sum_{i=1}^{N} k_i^{(\mathrm{in})} = \sum_{\ell=1}^{N}\sum_{m=1}^{N} T_{\ell m} = \mathscr{T}, \tag{3}$$

where $\mathscr{T}$ is the total number of trips considered in the origin-destination matrix.

In Fig. 3, we show the origin-destination matrix and the respective matrix of distances $\mathbf{D}$ obtained from taxi trips in 2015. The resulting matrices incorporate the flow of vehicles between $N = 4353$ high demand zones.

Let us now analyze the statistical properties of the directed weighted network associated with mobility in New York City. In order to do so, we show in Fig. 4 two probability distributions: one associated to the in-degree of the network $k_i^{(\mathrm{in})}$ (Fig. 4(a)) and the other one associated with the out-degree of the network $k_i^{(\mathrm{out})}$ (Fig. 4(b)). We explore all the in and out-degrees, for seven years, from 2009 to 2015, in the interval $10^3 \leq k \leq 10^6$. With the aim of finding the best fit of the aggregated data of mobility for these distributions, we used the tools and procedures described by Clauset *et al.* (2009) as given in ref. [37], that are implemented in the *powerlaw* package library described in references[38–40]. In order to decide the best fit and perform a proper statistical analysis, we explore several candidates for the distribution models: power law, power law with an exponential cutoff, exponential, stretched exponential and log-normal.

For the statistical distributions considered, we calculate the Kolmogorov-Smirnov (KS) distance between them in a pairwise fashion. This KS distance gives us a first indicator (goodness of fit) of the proximity of the data and the proposed distribution model. Then, we compare the different distributions via a likelihood ratio test by calculating the log-likelihood function of each one of the selected distributions. The sign of this ratio gives us a criterion to discriminate between distributions (see reference[37]). After this model selection, the best two fits were the power law with an exponential cutoff (EC), with a probability density[37]:

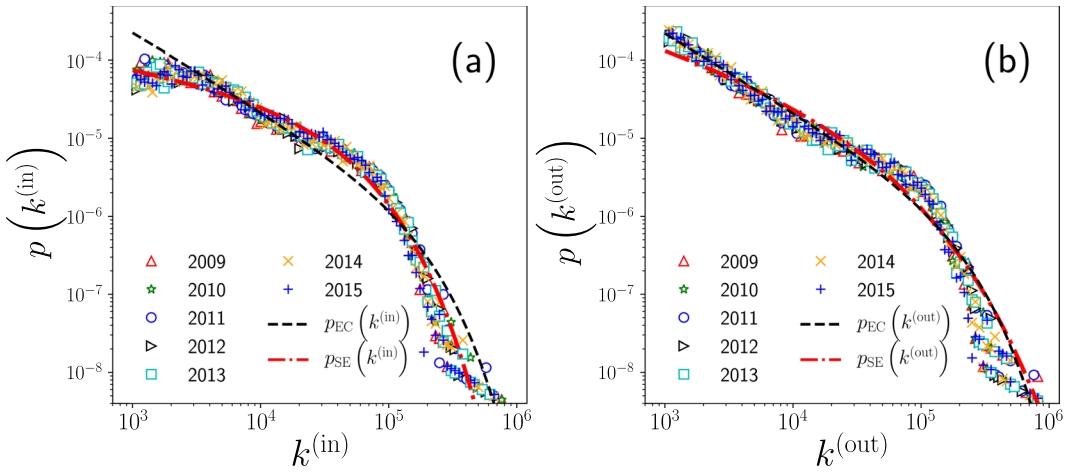

Figure 4. Statistical analysis of the number of taxi trips that depart and arrive in high demand zones in New York City. We present the probability density for the values of the degrees: (**a**) $k_i^{(\text{in})}$ and (**b**) $k_i^{(\text{out})}$ defined in Eqs. 1–2 for $i = 1, 2, \ldots, N$, where N is the number of high activity zones presented in Table 1 for each of the years explored. The results were obtained with normalized counts using logarithmically spaced bins. In both cases, we show with different curves, the power law with an exponential cutoff $p_{\text{EC}}(k)$ in Eq. 4 and the stretched exponential fit $p_{\text{SE}}(k)$ in Eq. 5.

$$p_{\text{EC}}(k) = \frac{\lambda^{1-\gamma}}{\Gamma(1-\gamma, \lambda k_{\text{min}})} k^{-\gamma} e^{-\lambda k} \tag{4}$$

and the stretched exponential (SE)[37]:

$$p_{\text{SE}}(k) = \beta\lambda k^{\beta-1} e^{-\lambda(k^{\beta}-k_{\text{min}}^{\beta})} \tag{5}$$

where k represents the degree, k_{min} is the minimum value considered in the fit and $\Gamma(x, y)$ denotes the incomplete gamma function. Notice that both distributions have two parameters, that we will distinguish with a superindex EC for the power law with an exponential cutoff, and with a superindex SE for the stretched exponential; however, we will not indicate these superindexes in Eqs. 4–5 to simplify the notation. We use λ^{EC} and γ^{EC} for the power law with an exponential cutoff and λ^{SE} and β^{SE} for the stretched exponential.

For the in-degrees in Fig. 4(a), the best fit is the stretched exponential with parameters $\beta_{\text{in}}^{\text{SE}} = 0.708$ and $\lambda_{\text{in}}^{\text{SE}} = 4.138 \times 10^{-5}$; in a similar way, for the power law with exponential cutoff $\gamma_{\text{in}}^{\text{EC}} = 1.00000000041$ and $\lambda_{\text{in}}^{\text{EC}} = 6.730 \times 10^{-6}$. On the other hand, the same analysis for the out-degrees in Fig. 4(b) concludes that the best fit is the power law with an exponential cutoff with parameters $\gamma_{\text{out}}^{\text{EC}} = 1.0000000025$ and $\lambda_{\text{out}}^{\text{EC}} = 6.086 \times 10^{-6}$; in addition, for the stretched exponential $\beta_{\text{out}}^{\text{SE}} = 0.495$ and $\lambda_{\text{out}}^{\text{SE}} = 6.834 \times 10^{-5}$.

It is surprising that both exponents $\gamma_{\text{in}}^{\text{EC}}$ and $\gamma_{\text{out}}^{\text{EC}}$ are extremely close to the value one. Thus, both distributions are well described by the power law $p(k) \propto k^{-1}$ in some range of in and out degrees.

Transition probabilities. All the information in the origin-destination matrix and in the degrees $k_i^{(\text{in})}$ and $k_i^{(\text{out})}$ allow us to analyze and understand the spatial activity of taxis as a dynamical process in a spatial directed weighted network. In this way, we can describe statistically the global dynamics of taxis in terms of transition probabilities between high demand zones of this service.

The transition probability $w_{i\to j}^{(\text{out})}$ between zones i and j is defined in terms of the origin-destination matrix as:

$$w_{i\to j}^{(\text{out})} = \frac{T_{ij}}{k_i^{(\text{out})}}. \tag{6}$$

With this definition, the transition probabilities satisfy the normalization condition:

$$\sum_{\ell=1}^{N} w_{i\to \ell}^{(\text{out})} = 1. \tag{7}$$

With the transition probabilities $w_{i\to j}^{(\text{out})}$, we can explore the relationship between the information in the origin-destination matrix and the matrix of distances; these matrices were presented in Fig. 3. Now, to study this connection, we calculate $w_{i\to j}^{(\text{out})}$ by using the definition in Eq. 6; for each value, we have the corresponding geographical distance d_{ij} between i and j as an entry in the distance matrix $\mathbf{D}$.

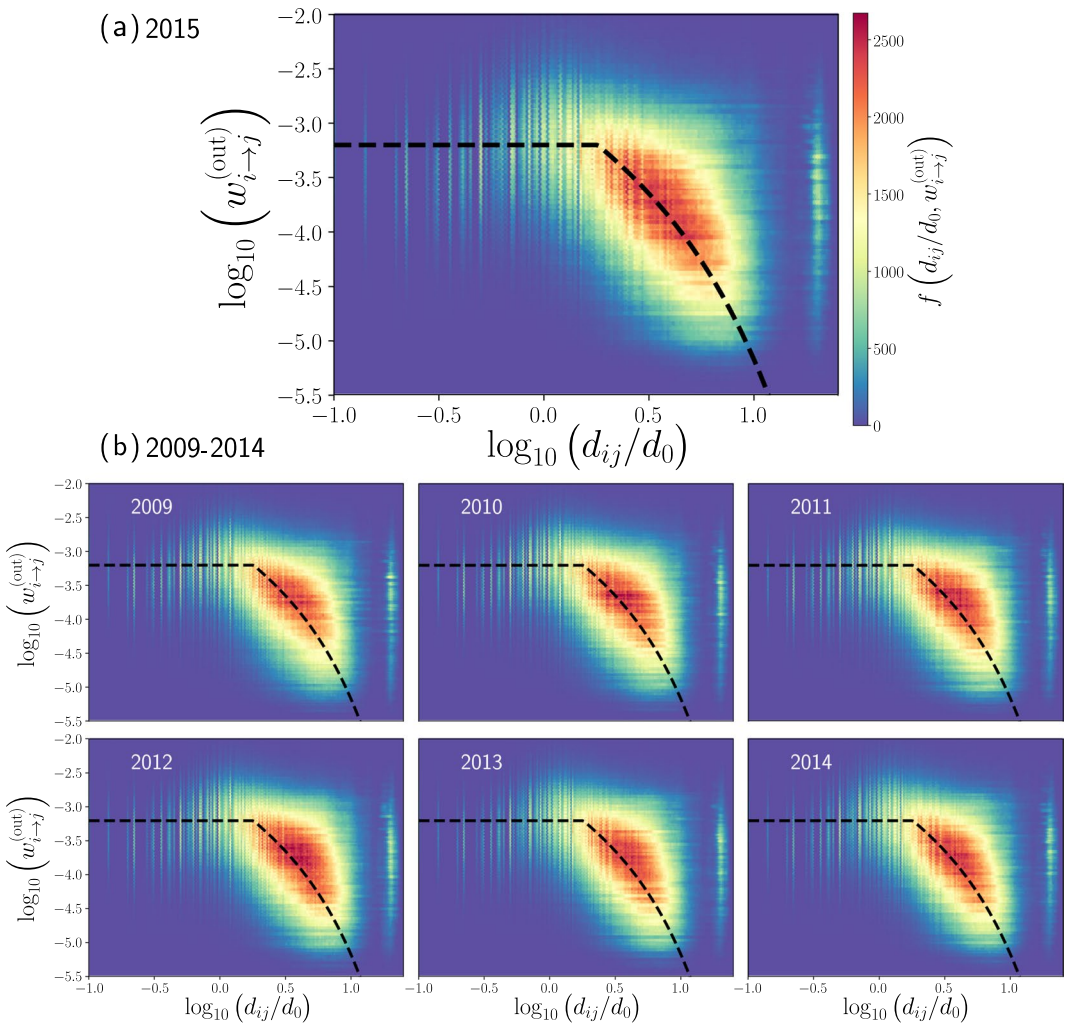

Figure 5. Transition probabilities between zones of taxi trips in New York City. In (**a**) we present the results obtained for the year 2015 with origin-destination matrix and the respective distances presented in Fig. 3. In (**b**) we depict our findings for each year from 2009 to 2014. In all these cases, we analyze the non-null transition probabilities $w_{i\to j}^{(\mathrm{out})}$ and the geographical distance d_{ij} between zones i and j. We show hexagonally binned two-dimensional histograms for the logarithm of $w_{i\to j}^{(\mathrm{out})}$ and the logarithm of d_{ij}/d_0 where $d_0 = 1$ Km is a reference distance. The values codified in the colorbar represent the frequencies denoted as $f\left(d_{ij}/d_0,\, w_{i\to j}^{(\mathrm{out})}\right)$ of the pairs $\left(\log_{10}\left(\frac{d_{ij}}{d_0}\right),\, \log_{10} w_{i\to j}^{(\mathrm{out})}\right)$ found in each hexagonal bin. Dashed lines are used as a guide and represent the behavior $w_{i\to j}^{(\mathrm{out})}$ constant, for $d \leq 1.8$ Km, and $w_{i\to j}^{(\mathrm{out})} \propto d_{ij}^{-1} e^{-\beta\left(d_{ij}-R\right)}$ with $\beta = 0.15$ Km^{-1} for $d > 1.8$ Km.

In Fig. 5, we depict the logarithm of the transition probability $w_{i\to j}^{(\mathrm{out})}$ as a function of the logarithm of the relation d_{ij}/d_0 where $d_0 = 1$ Km is a reference length. In Fig. 5(a), we consider all the non-null transition probabilities $w_{i\to j}^{(\mathrm{out})}$ and distances d_{ij}, for the annual data records of the taxi's activity in 2015; we obtain a distribution of points $\left(\log_{10}\left(\frac{d_{ij}}{d_0}\right),\, \log_{10} w_{i\to j}^{(\mathrm{out})}\right)$ for all the zones with high demand ($i, j = 1, 2, \ldots, N$). We show the results as a two-dimensional histogram that quantifies the frequencies of these values in hexagonal bin counts.

Our findings in Fig. 5 reveal that the transition probabilities of taxis are approximately constant $w_{i\to j}^{(\mathrm{out})} = 10^c$ for distances less than a characteristic value $R = 1.8$ Km. In contrast, for distances greater than R, the transition probabilities are well described by a power law with an exponential cutoff relation:

$$w_{i\to j}^{(\mathrm{out})} = a\frac{R}{d_{ij}} e^{-\beta\left(d_{ij}-R\right)} \qquad \text{for} \qquad d_{ij} > R, \tag{8}$$

where continuity for $d_{ij} = R$ requires $a = 10^c$. Now, to find the best fit, we analyze the pairs $\left(\log_{10}\left(\frac{d_{ij}}{d_0}\right),\, \log_{10} w_{i\to j}^{(\mathrm{out})}\right)$ presented in Fig. 5 for values 0.1 Km $\leq d_{ij} \leq$ 11 Km. We divide the data considering pairs in the region $d_{ij} \leq R$ and $d_{ij} > R$ with $R = 1.8$ Km. In Fig. 6(a) we show the statistical analysis of the values $c = \log_{10} w_{i\to j}^{(\mathrm{out})}$ found for $d_{ij} \leq R$, we see that the values c are distributed with a pronounced peak around $c = -3.2$, we use this value to describe the

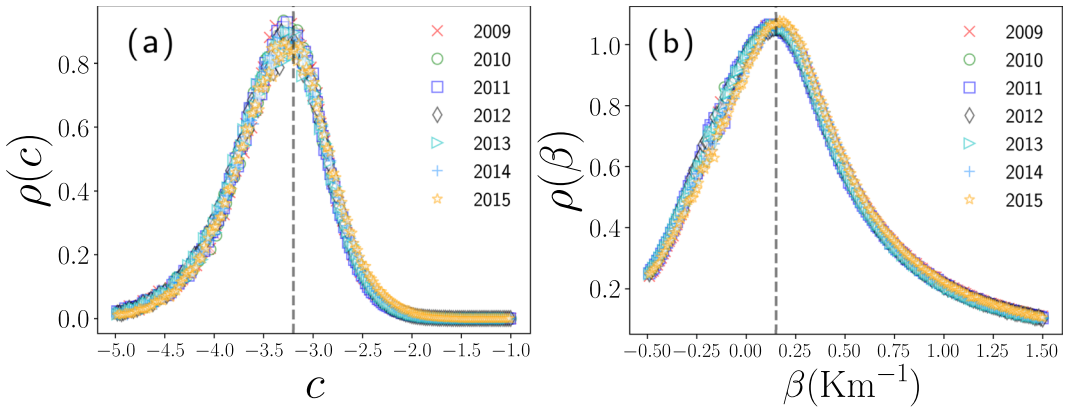

Figure 6. Statistical analysis of the parameters c and β. We present the probability density ρ of the numerical values c and β found for each pair $\left(\log_{10}\left(\frac{d_{ij}}{d_0}\right), \log_{10} w_{i\to j}^{(\mathrm{out})}\right)$ in the interval 0.1 Km $\leq d_{ij} \leq$ 11 Km for the years 2009, 2010, …, 2015. (**a**) Values $c = \log_{10} w_{i\to j}^{(\mathrm{out})}$ for $d_{ij} \leq R = 1.8$ Km, (**b**) values β obtained from Eq. 8 for $d_{ij} > R$. Vertical dashed lines represent the values $c = -3.2$ and $\beta = 0.15$ Km^{-1}.

probabilities of transition. In a similar way, once defined c, we calculate β in Eq. 8 for $d_{ij} > R$. In Fig. 6(b), we analyze the probability density of the values β and we identify a peak around $\beta = 0.15$ Km^{-1}.

The piecewise approximations described by the values $R = 1.8$ Km, $c = -3.2$ and $\beta = 0.15$ Km^{-1} are represented with dashed lines in Fig. 5. A similar behavior has been detected in the analysis of the transportation network of stations in bicycle sharing systems operating in New York City and Chicago[21]. In these cases, the value $R \approx 1$ Km defines local displacements and the long-range dynamics is well described by $w_{i\to j}^{(\mathrm{out})} \propto d_{ij}^{-2}$. In this way, in bike-sharing systems R is reduced in comparison to our findings for taxi trips; in addition, the long-range spatial activity qualitatively has similar characteristics to those observed in Fig. 5.

OD rank. The transition matrix $\mathbf{W}^{(\mathrm{out})}$ with elements $w_{i\to j}^{(\mathrm{out})}$ defined in Eq. 6 allow us to understand human mobility as a dynamical process in a spatial directed weighted network. Well-known results in stochastic processes apply for the transition matrix $\mathbf{W}^{(\mathrm{out})}$[10]. In most cases, origin-destination matrices are non-symmetric; as a consequence, it is convenient to analyze the transition matrix $\mathbf{W}^{(\mathrm{out})}$ establishing an analogy with the Google matrix[41], with a mathematical structure entirely general that applies to any graph or network in any domain[42]. In the following, we explore how by using this connection, the eigenvalues and eigenvectors of $\mathbf{W}^{(\mathrm{out})}$ give valuable information to understand the movement of taxis.

The transition matrix $\mathbf{W}^{(\mathrm{out})}$ has left and right eigenvectors. Left eigenvectors $\overrightarrow{\Phi}_j$ with elements $\phi_j(i)$ satisfy:

$$\overrightarrow{\Phi}_j \mathbf{W}^{(\mathrm{out})} = \lambda_j \overrightarrow{\Phi}_j \qquad \text{for} \qquad j = 1, 2, \ldots, N, \tag{9}$$

where $\{\lambda_j\}_{j=1}^{N}$ are the eigenvalues of the transition matrix. Right and left eigenvectors form an orthonormal base and have the same eigenvalues. On the other hand, the stochastic matrix $\mathbf{W}^{(\mathrm{out})}$ fulfills Eq. 7 and, by definition, the elements of T_{ij} satisfy $T_{ij} \geq 0$; therefore, $\mathbf{W}^{(\mathrm{out})}$ belongs to the class of Perron-Frobenius operators with a possibly degenerate unit eigenvalue $\lambda = 1$ and other eigenvalues obeying $|\lambda| \leq 1$ (see[43] for details).

In Fig. 7(a) we plot the eigenvalues of the transition matrix $\mathbf{W}^{(\mathrm{out})}$ for taxi trips in New York City in 2015. We use the origin-destination matrix in Fig. 3(a) and the definition in Eq. 6. The results were obtained numerically and, due to the asymmetry of the origin-destination matrix, the eigenvalues are complex numbers. In Fig. 7(a) we show the real and imaginary part of each of the eigenvalues λ_i for $i = 1, 2, \ldots, N = 4\,353$. In this analysis, we found that only one eigenvalue satisfies $\lambda = 1$, a result that reveals that the directed network associated with the mobility between sites of high demand for taxis is connected. Therefore, the links in the network connect all the zones. This particular result can be interpreted using the terminology of random walks on networks. In this case, the movement of a random walker defined in terms of the transition matrix $\mathbf{W}^{(\mathrm{out})}$ is capable to visit any node of the network only by moving on the links, independently of the initial configuration. As we mentioned before, the high connectivity observed in the origin-destination matrix is a consequence of considering high demand zones with a criterion that requires a high number of departures and arrivals in each zone avoiding the emergence of isolated parts. However, the approach developed is general and in other cases, similar spectral analysis of the transition matrix could be an important tool to identify disconnected parts in a transportation system.

In addition to the eigenvalues, the respective eigenvectors of the transition matrix provide valuable information about dynamical processes on networks[10,19]. In particular, the left eigenvector associated with the eigenvalue $\lambda = 1$ defines a ranking vector $\overrightarrow{\mathbf{P}}^{\infty}$ with elements P_i^{∞} for $i = 1, 2, \ldots, N$ and satisfies $\overrightarrow{\mathbf{P}}^{\infty} \mathbf{W}^{(\mathrm{out})} = \overrightarrow{\mathbf{P}}^{\infty}$, therefore:

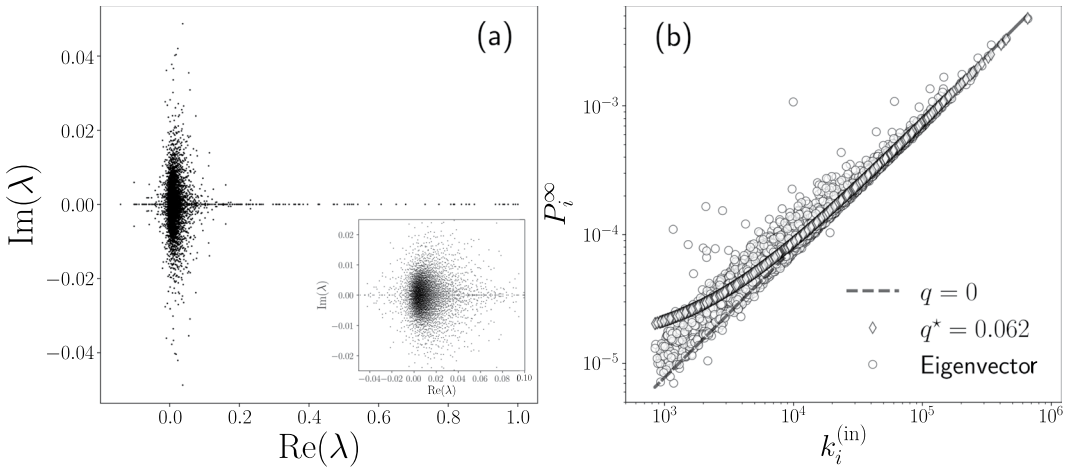

Figure 7. Numerical analysis of the eigenvalues and OD rank of the transition matrix $\mathbf{W}^{(\mathrm{out})}$. We analyze the transition probability matrix for the taxi's flow in 2015 with origin-destination matrix presented in Fig. 3 with $N = 4\,353$ high demand zones. In (**a**), we show the eigenvalues λ of $\mathbf{W}^{(\mathrm{out})}$ satisfying Eq. 9. In this way, we have 4 353 values represented in the complex plane with dots; in the inset, we depict the results for the eigenvalues in a region close to the origin, where we observe more eigenvalues with a non-null complex part. In (**b**) we plot the components P_i^∞ of the eigenvector $\overrightarrow{\mathbf{P}}^\infty$ with eigenvalue $\lambda = 1$; we represent the numerical values of P_i^∞ in terms of the respective degree $k_i^{(\mathrm{in})}$ for $i = 1, 2, \ldots, N$. We also show the values $P_i^\infty(q)$ obtained with Eq. 11 for $q = 0$ and the best fit $q^* = 0.062$.

$$\sum_{\ell=1}^{N} P_\ell^\infty w_{\ell\to j}^{(\mathrm{out})} = P_j^\infty. \tag{10}$$

In the study of random walks on networks, the vector $\overrightarrow{\mathbf{P}}^\infty$ is the stationary probability distribution. The value P_i^∞ gives the probability of a random walker to reach the node i after a large number of steps[19]. In the context of the Google matrix, the vector $\overrightarrow{\mathbf{P}}^\infty$ determines the importance of a node in a network establishing a PageRank of the Web[43]. In the analysis of mobility with a transition matrix $\mathbf{W}^{(\mathrm{out})}$, the vector $\overrightarrow{\mathbf{P}}^\infty$ defines a ranking of the zones used in the definition of the origin-destination matrix. Due to this connection, we call this ranking "*OD rank*".

In Fig. 7(b), we show the results obtained numerically for the OD rank $\overrightarrow{\mathbf{P}}^\infty$ associated with the eigenvalue $\lambda = 1$ of the transition matrix $\mathbf{W}^{(\mathrm{out})}$ that describes the taxi's flow in 2015. Our findings in this figure reveal a connection between the OD rank P_i^∞ of a zone i and the respective in-degree $k_i^{(\mathrm{in})}$. In a similar way to the findings for the PageRank algorithm for Google, the stationary probability distribution $\overrightarrow{\mathbf{P}}^\infty$ is a measure of the popularity of nodes that is mostly due to the in-degree dependence; in a mean-field approximation the stationary distribution of the PageRank algorithm is given by[44]:

$$P_i^\infty(q) = \frac{q}{N} + (1-q)\frac{k_i^{(\mathrm{in})}}{\mathcal{T}}, \tag{11}$$

where $0 \le q \le 1$. Searching the optimal value q^* that minimizes the quadratic error $S(q) = \sum_{\ell=1}^{N}(P_\ell^\infty - P_\ell^\infty(q))^2$, we get for the best fit:

$$q^\star = \frac{\sum_{\ell=1}^{N}(p_\ell)^2 - \sum_{\ell=1}^{N} P_\ell^\infty p_\ell}{\sum_{\ell=1}^{N}(p_\ell)^2 - \frac{1}{N}} \qquad \text{with} \qquad p_\ell = \frac{k_\ell^{(\mathrm{in})}}{\mathcal{T}}. \tag{12}$$

In Fig. 7(b) we illustrate the approximation given by Eq. 11, for $q = 0$ and $q^* = 0.062$, obtained for the best fit. However, Eq. 11 is a mean field result and important deviations may appear[10,44–46]. The result given by Eq. 11 makes sense in the description of taxis since the importance of a high demand zone can be defined in terms of the number of taxi trips $k^{(\mathrm{in})}$ that arrive at this specific location. For example, in our schematic illustration presented in Fig. 1(a), now we understand that the bars with the value $k^{(\mathrm{in})}$ determine the importance of the zones.

The transition probability matrix $\mathbf{W}^{(\mathrm{out})}$ defined in Eq. 6 captures all the information about the system's global activity. We think that an OD rank of the zones defined as $\overrightarrow{\mathbf{P}}^\infty$ can be a valuable measure in the analysis of different transportation systems and a complement to other types of ranking algorithms introduced to determine location attractiveness incorporating geographic considerations into the PageRank algorithm[47–49].

Random walk strategy. The results obtained before for the relationship of the transition probabilities describing the flow of taxis between zones and the geographical distances separating these locations, suggest that the spatial dynamics can be approximately described by a model with constant transitions to zones in a local

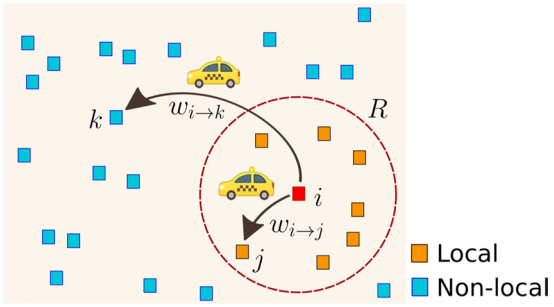

Figure 8. A schematic illustration of the mobility of taxis between high demand zones. There are two types of trips from a particular location *i*: First, to a site *j* inside a circular region of radius *R* centered in the location *i*, the probabilities to have a trip to these zones are constant; and, second, a trip to a zone *k* outside the circle of radius *R*. In this case, the probability to have this long-range movement decays as a power law with an exponential cutoff proportional to $e^{-\beta(d_{ik}-R)}d_{ik}^{-1}$, where d_{ik} is the geographical distance between *i* and *k*.

neighborhood within a distance *R*, and a long-range dynamics defined by probabilities of transition proportional to $e^{-\beta(d_{ij}-R)}d_{ij}^{-1}$. The analysis of more than a billion trips reveals a particular emergent pattern in the spatial activity. The movement of taxis between high demand zones can be classified into two types of trips with particular characteristics illustrated in Fig. 8. We have local displacements for which a taxi departs from a high demand site and the probability of moving to another site of high activity is independent of the distance that separates them if they are located at a distance less than a value *R*. On the other hand, there may also be long-range displacements for which the separation between origin and destinations require distances greater than *R*. For this type of movements, we find that the probability of having a long-range trip depends on the distance and these particular transitions have characteristics observed in truncated Lévy flights.

In this way, to describe the global activity of the taxi's mobility we use the model:

$$w_{i\to j}^{(\text{model})}(R,\,\beta) = \frac{\Omega_{ij}(R,\,\beta)}{\sum_{\ell=1}^{N}\Omega_{i\ell}(R,\,\beta)}, \tag{13}$$

where:

$$\Omega_{ij}(R,\,\beta) = \begin{cases} 1 & \text{for} \quad 0 \le d_{ij} \le R, \\ \left(R/d_{ij}\right)e^{-\beta\left(d_{ij}-R\right)} & \text{for} \quad R < d_{ij}. \end{cases} \tag{14}$$

In this model, β and *R* are positive real parameters. The transition probabilities defined in Eqs. 13 and 14 are illustrated in Fig. 8. The radius *R* determines a neighborhood around each zone where the trips occur with equal probability to move from the initial site to any of the high demand zones in this region. Therefore, the displacements are independent of the geographical distance between origin and destination. That is, if there are *S* sites inside a circle of radius *R*, the probability of going to any of these sites is uniform. Additionally, for places beyond the local neighborhood, for distances greater than *R*, the transition probability decays as a power law with an exponential cutoff of the distance and is proportional to $e^{-\beta(d_{ij}-R)}d_{ij}^{-1}$. In this way, the parameter *R* defines a characteristic length of the local neighborhood and β controls the probability to have long-range displacements. In particular, in the limit $\beta \to \infty$ the dynamics becomes local. We introduced a similar model with long-range transitions proportional to $d_{ij}^{-\alpha}$ ($\alpha > 0$) in reference[8] in the context of human mobility and encounter networks. In this case, the resulting dynamics can be similar to a rank model[50–52] and a gravity model[3,53–55]. It is worth mentioning that the inverse of the parameter β in Eq. 14 gives us a characteristic distance; this exponential cutoff takes into account the finite size effect associated with a finite system like New York City.

In our previous analysis in Fig. 5, we found that $R \approx 1.8$ Km. This value defines what we understand as a local neighborhood for this transport system. On the other hand, for distances $d_{ij} > R$, the probability to have a trip to a zone is highly influenced by the geographical distance and this long-range dynamics is determined by the values $e^{-\beta(d_{ij}-R)}d_{ij}^{-1}$ with $\beta = 0.15$ Km^{-1}.

In the following part, we explore the predictions of this model for the annual global activity of taxi displacements in New York City by using the parameters $R = 1.8$ Km and $\beta = 0.15$ Km^{-1} found in the analysis of the taxi's flow between high demand zones. In addition to Eqs. 13 and 14, that model the displacement between high demand zones, it is important to consider that these zones have different relevance in the whole dynamics, i.e., a trip can start from different zones with non-uniform probabilities. This fact is well described by the values of the out-degree $k_i^{(\text{out})}$ defined in Eq. 2 that gives the number of trips with origin in the zone *i*. In addition, from the results in Fig. 4, we know that the values $k_i^{(\text{out})}$ follow a hierarchical distribution with probabilities that decay as $p(k) \propto k^{-\gamma}e^{-\lambda k}$ where *k* represent the values of the out-degree. This result is observed in the annual datasets from 2009-2015. In this way, we simulate the dynamics of multiple taxis that start from an initial zone chosen randomly with a probability proportional to the values $\{k_i^{(\text{out})}\}_{i=1}^{N}$ that quantify the importance of each zone in the city. Then,

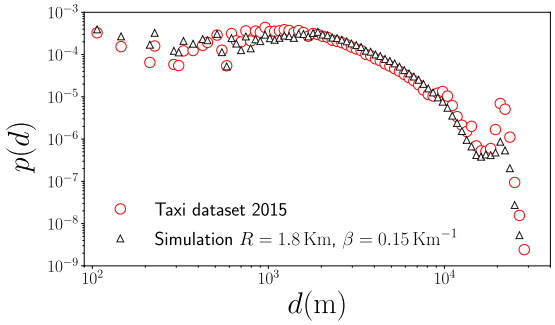

Figure 9. Statistical analysis of displacements of taxi trips in New York City. We depict the probability density $p(d)$ of the geographical distance d between the departure zone and the final destination of taxis. We present statistics obtained from the analysis of the complete dataset for displacements in 2015 and data generated by using Monte Carlo simulations with transition probabilities $w_{i\to j}^{(\mathrm{model})}(R,\ \beta)$ defined by the our model in Eqs. 13 and 14 with $R = 1.8$ Km and $\beta = 0.15$ Km^{-1}. In both cases we use logarithmic spaced bin counts for distances between $10^2\ \mathrm{m} \leq d \leq 4 \times 10^4\ \mathrm{m}$.

a displacement is generated randomly from the origin site to a final zone by using the transition probabilities in Eq. 13; this algorithm is repeated to generate, through Monte Carlo simulations, the same number of displacements between high demand zones, as reported in Table 1.

In Fig. 9 we present the statistical analysis of the taxi's displacements d generated randomly and the real values considering the activity in New York City in 2015. In our simulation, we generate 128 984 657 random displacements following the model in Eq. 13, with $\beta = 0.15$ Km^{-1} and $R = 1.8$ Km. Our findings show an agreement between the predictions of the model and the real dynamics. However, we observe that the predictions do not agree with the real data around $d = 10$ Km and $d = 20$ Km; this is a consequence of the singular dynamics induced by the two airports in New York City. An accurate modeling capturing the effects of these very attractive sites in a city requires modifications to the model explored. This fact is also visible in Fig. 5 where, for distances around 20 Km, we see different values of the transition probability that are not described by a model with long-range trips following $w_{i\to j}^{(\mathrm{out})} \propto e^{-\beta(d_{ij}-R)} d_{ij}^{-1}$.

Finally, we repeat the same procedure to compare the predictions of the model with respect to the real data for taxi's activity from 2009 to 2014. Our results for Monte Carlo simulations are presented in Fig. 10. We observe the same characteristics found in Fig. 9, with a good agreement between model and the data. The number of locations of high demand N and the number of displacements analyzed for each year are reported in detail in Table 1. The results in Table 1 also reveal that in average, in a year, approximately 43% of the trips are local movements for which the geographical distance $d \leq 1.8$ Km, the rest of the trips are non-local with $d > 1.8$ Km.

Discussion

In this research, we explore the massive records of more than one billion taxi-trips in New York City from January 2009 to December 2015. With this dataset of seven years, we generate an origin-destination matrix that has detail information of a vast number of trips. The mobility in New York City can be described as a directed weighted network that connects different zones of high demand for taxis. Each zone is characterized by the number of trips that arrive or depart from it and corresponds to nodes in the network. The arrivals and departures are the in-degrees and out-degrees of the directed network, and the flow gives different weights to the links of this spatial network.

We present a statistical analysis of the travel distance of each trip and found a long-range distribution that is almost the same for each of the seven years studied. On the other hand, the degree distributions, for the in and out degrees are, respectively, well modeled by a stretched exponential and a power law with an exponential cutoff. By defining the transition probabilities between zones, given by the origin-destination matrix and the out-degree, we are able to obtain a rank, called "*OD rank*", analogous to the page rank of Google. We calculate the spectrum of eigenvalues and the main eigenvector, which is related to the in degree. The components of this eigenvector give the more relevant and attractive places in New York City, in terms of taxi trips.

The dependence of the transition probabilities with the distance between zones is obtained from the dataset, and based on that, we introduce a model that captures the global dynamics of trips. The data and the model describe, for short distances, a local dynamics independent of the spatial distance, and, for large distances, a dynamics that decays with distance as a power law with an exponential cutoff. The data agrees quantitatively with Monte Carlo simulations based on our model.

Finally, considering the taxi trips as a proxy of human mobility in cities, it might be possible that the long-range mobility and other features found for New York City would be rather general, and thus we expect a similar behavior in other large cities around the world for which these ideas can be applied as well.

Methods

Dataset description. In this section, we present a global description of the records explored to study the spatial dynamics in New York City. We use data for the activity of taxi trips from January 2009 to December 2015; these datasets are available to the public by the Taxi and Limousine Commission in the New York City open data website[56]. The data available include information for all taxi trips in New York City when the taxis are in service.

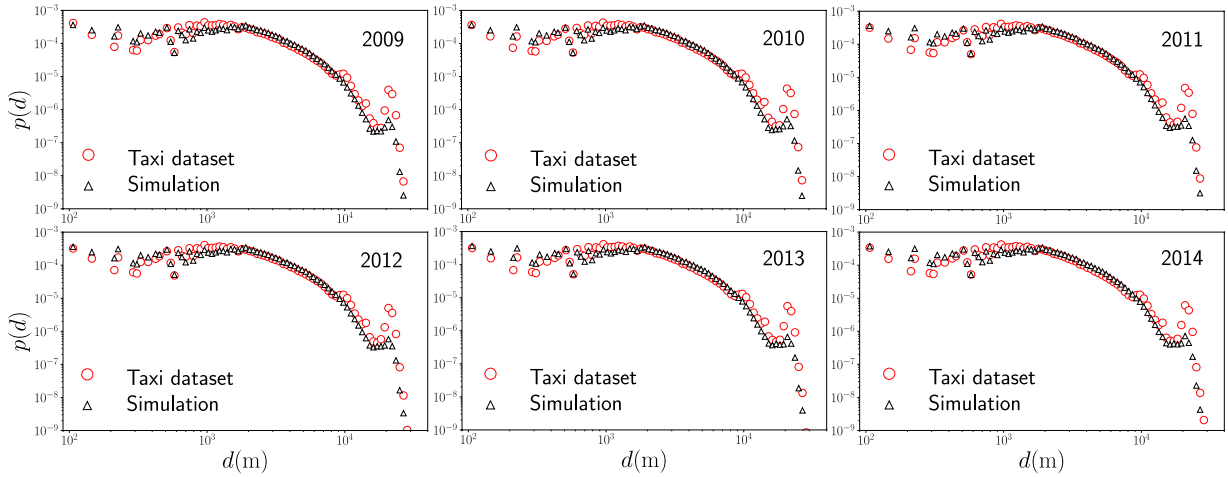

Figure 10. Probability density $p(d)$ of the geographical distance d between the departure zone and final destination of taxis and the results generated through Monte Carlo simulations with transition probabilities between high demand zones $w_{i\to j}^{(\mathrm{model})}(R,\ \beta)$ with $R = 1.8$ Km and $\beta = 0.15$ Km^{-1}.

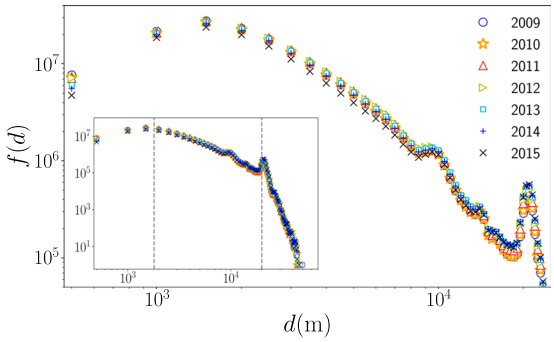

Figure 11. Statistics of displacements d of taxi trips in New York City. We depict the frequency $f(d)$ of the geographical distance d between the origin and destination of taxis. The results are obtained from annual datasets between January 2009 to December 2015. In the inset, we present $f(d)$ as a function of d for the analysis of all the distances with a scale in the frequencies that ranges from 10^0 to 10^8. The two vertical dashed lines represent $d = 1.8$ Km and $d = 20$ Km. Additional information about the datasets explored is presented in Table 2.

Year	$\mathscr{T}$	$\langle d\rangle$(Km)	d_{max}(Km)	% $0 \le d < 1.8$ Km	% 1.8 Km $\le d <$ 20 Km	% $d \ge 20$ Km
2009	167 165 746	3.14	49.02	42.76	56.32	0.92
2010	163 913 012	3.19	51.87	42.05	56.96	0.99
2011	171 166 041	3.27	45.16	41.05	57.85	1.1
2012	173 087 239	3.34	44.91	40.37	58.47	1.16
2013	168 937 296	3.36	47.67	40.45	58.28	1.27
2014	160 822 602	3.38	43.78	40.92	57.72	1.36
2015	142 958 901	3.41	45.95	41.82	56.62	1.56
2009–2015	1 148 050 837	3.30	51.87	41.33	57.49	1.18

Table 2. Taxi records and displacements in New York City. We analyze taxi trips records from January 2009 to December 2015. Here, $\mathscr{T}$ is the total number of trips, the length $\langle d\rangle$ is the average distance between the initial and final location whereas $d_{\max}$ is the length of the maximum displacement observed in each dataset. On the other hand, in the last three columns we include the fraction of displacements (as percentages) in the intervals $0 \le d < 1.8$ Km, 1.8 Km $\le d < 20$ Km and for values of d larger than 20 Km. In the last row, we present the results obtained for the whole dataset.

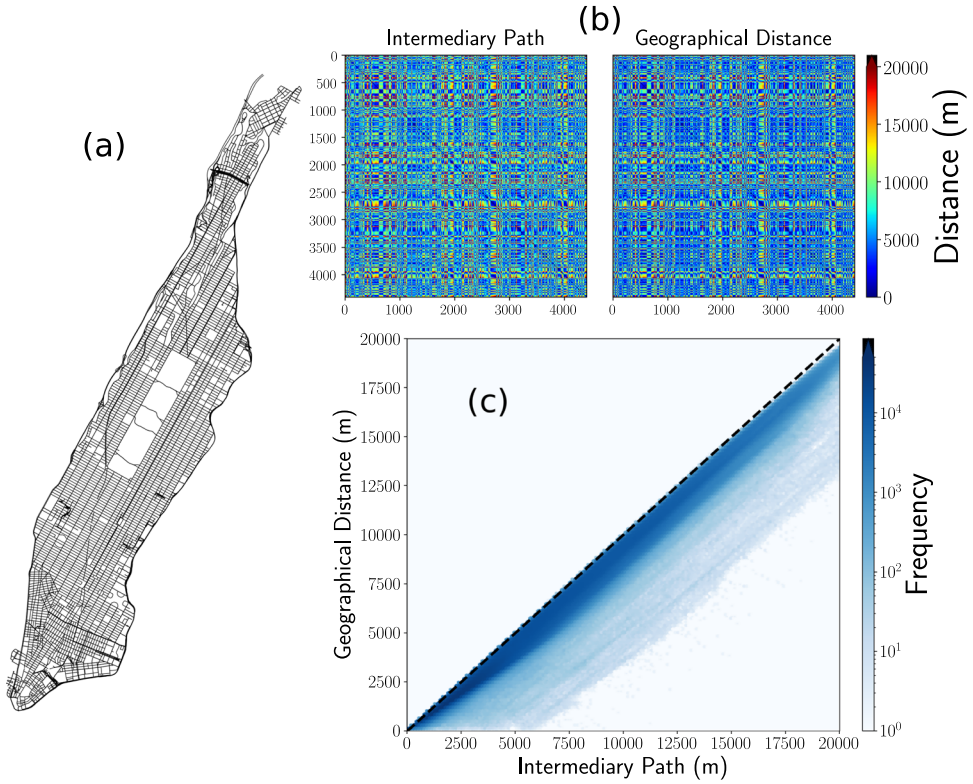

Figure 12. Distances between intersections in Manhattan. (**a**) Manhattan's street network, (**b**) Distance matrices for 4 409 intersections in this network. We depict the results for the length of the intermediary path and the geographical distance between these intersections; the distances are indicated by the colorbar. In (**c**) we present the hexagonal bin counts for the geographical distances and the respective length of the intermediary path. We depict a dashed line, with unit slope, that represents the case when the two distances are the same. Clearly, since the intermediary path is always greater or equal than the geographic distance, we only have data in the lower triangle of the figure. We show with a colorbar the frequencies for the values found in each bin. The street map, intersections, and intermediary paths were obtained and analyzed with the OSMnx package[57,58].

The records comprise several fields capturing pick-up and drop-off dates and times, pick-up (origin) and drop-off (destination) locations, itemized fares, rate types, payment types and driver-reported passenger counts[56].

Now, to complement the information in Fig. 2, and identify other global characteristics in the taxi's spatial activity, we analyze the geographical distance d between the origin and destinations in each trip calculated from the longitude and latitude coordinates of these locations reported in the database. Here, it is worth mentioning that other types of distances can be implemented; in particular, the distance of the path in the street network connecting origin and destinations. In fact, powerful techniques have been introduced exploring taxi trips in New York City to estimate the driving distance based on the origin and destination coordinates[59]. However, due to the different paths that a taxi can follow to carry out each trip, in the following we use the geographical distance d. In Fig. 11, we present the statistical analysis of the geographical distances d. We depict the frequencies $f(d)$ of the displacements obtained from uniform bin counts with $\Delta d = 500$ m for taxi trips. Different markers show the results for the analysis in a year. We can see that the frequencies $f(d)$ maintain the same characteristics from 2009 to 2015, and the statistics reveal three important intervals: the first for $d < 1.8$ Km with higher values of the frequencies, a second interval for 1.8 Km $\leq d < 20$ Km where $f(d)$ gradually decays and finally, for distances around 20 Km, we identify a peak that decays rapidly with the distance; this peak is associated with large displacements from Manhattan to the JFK airport (as a reference, the geographical distance between Times Square and the JFK airport is 20.6 Km). In a similar way, we identify another relative maximum at $d = 10$ Km: this increase in the frequencies is associated with trips between Manhattan and La Guardia airport (with $d = 9.8$ Km between Times Square and this airport). These are examples of how important locations can induce long-range dynamics in the taxi's mobility. In this case, the two airports in New York City influence the taxi transportation mode in the whole city. This important feature has been observed in other cities with airports located at the city's periphery (a particular case is reported in[31]).

In Table 2, we summarize the global information found for the spatial dynamics per year. We present the number of taxi trips analyzed, the average distance, the largest distance traveled as well as the fraction of trips with distances at different intervals. From the information in this table, when we examine the complete records from 2009 to 2015, we observe that 41.33% of the taxi trips have displacements with d less than 1.8 Km, whereas a 57.49% of the trips involve long-range displacements in the interval 1.8 Km $\leq d < 20$ Km, and only a 1.18% of the trips have d greater than 20 Km. The average displacement of trips is $\langle d \rangle = 3.3$ Km and the maximum value

observed in the records is 51.87 Km. All these quantities give us a first characterization of the spatial activity of the taxi transportation mode.

Geographical and shortest path distances. In the analysis of the information described before, we use the geographical distance d between origin and destination. This election is based on the fact that we only know the geographical coordinates of origins and destinations for each trip. However, another important quantity to consider is the length of the intermediary path that the vehicle follows on the street network. The information of the street network, the length, and direction of each street and all the intersections, can be obtained from different sources like OpenStreetMap[60] or generated by using specialized algorithms (see for example[59]). In general, the length of the geographical distance is less or equal than the length of the shortest path between two points in a city. In Fig. 12 we explore this relation for all the intersections in Manhattan's street network. We analyze the information available in OpenStreetMap[60] and the OSMnx Python package[57,58] to generate the street network depicted in Fig. 12(a). From this structure, we obtain the geographical coordinates of 4 409 intersections. In Fig. 12(b) we calculate all the distances between these intersections, taking into account the length of the intermediary path, and the respective geographical distance. The results are presented as matrices for which the entry l, m represents the respective distance between intersections l and m. The two matrices are similar; however, the matrix with intermediary paths is asymmetric since it includes the directions of the streets. In Fig. 12(c) we explore the relation between the two distances by plotting all the values presented in Fig. 12(b). The results reveal that a high fraction of the values is close to a linear relation. Similar results apply for the whole city and, in this way, the main features of the global activity of taxis can be analyzed by using only geographical distances between origin and destination. However, in other contexts, a description of the complete path followed by the vehicle is necessary. See refs. [35,59] for a detailed discussion and models for taxi's mobility at the level of intermediary paths.

Received: 3 December 2019; Accepted: 7 February 2020;
Published online: 04 March 2020

References

1. Batty, M. *The New Science of Cities* (MIT Press, Cambridge, MA, 2013).
2. Barthélemy, M. *The Structure and Dynamics of Cities: Urban Data Analysis and Theoretical Modeling* (Cambridge University Press, 2016).
3. Barbosa, H. *et al*. Human mobility: Models and applications. *Phys. Rep.* **734**, 1–74 (2018).
4. Louail, T. *et al*. From mobile phone data to the spatial structure of cities. *Sci. Rep.* **4**, 5276 (2014).
5. Louail, T. *et al*. Uncovering the spatial structure of mobility networks. *Nat. Commun.* **6**, 6007 (2015).
6. Lee, M. & Holme, P. Relating land use and human intra-city mobility. *PLoS ONE* **10**, e0140152 (2015).
7. Riascos, A. P. Universal scaling of the distribution of land in urban areas. *Phys. Rev. E* **96**, 032302 (2017).
8. Riascos, A. P. & Mateos, J. L. Emergence of encounter networks due to human mobility. *PLoS ONE* **12**, e0184532 (2017).
9. Newman, M. E. J. *Networks: An Introduction* (Oxford University Press, Oxford, 2010).
10. Barrat, A., Barthélemy, M. & Vespignani, A. *Dynamical Processes on Complex Networks* (Cambridge University Press, Cambridge, 2008).
11. Barabási, A.-L. *Network science* (Cambridge University Press, Cambridge, 2016).
12. Gallotti, R. & Barthélemy, M. Anatomy and efficiency of urban multimodal mobility. *Sci. Rep.* **4**, 6911 (2014).
13. Aleta, A., Meloni, S. & Moreno, Y. A Multilayer perspective for the analysis of urban transportation systems. *Sci. Rep.* **7**, 44359 (2017).
14. Boyer, D. *et al*. Scale-free foraging by primates emerges from their interaction with a complex environment. *Proc. R. Soc. B* **273**, 1743–1750 (2006).
15. Riascos, A. P. & Mateos, J. L. Long-range navigation on complex networks using Lévy random walks. *Phys. Rev. E* **86**, 056110 (2012).
16. Riascos, A. P. & Mateos, J. L. Fractional dynamics on networks: Emergence of anomalous diffusion and Lévy flights. *Phys. Rev. E* **90**, 032809 (2014).
17. de Nigris, S., Carletti, T. & Lambiotte, R. Onset of anomalous diffusion from local motion rules. *Phys. Rev. E* **95**, 022113 (2017).
18. Michelitsch, T. M., Collet, B. A., Riascos, A. P., Nowakowski, A. F. & Nicolleau, F. C. G. A. Fractional random walk lattice dynamics. *J. Phys. A: Math. Theor* **50**, 055003 (2017).
19. Michelitsch, T. M., Riascos, A. P., Collet, B. A., Nowakowski, A. F. & Nicolleau, F. C. G. A. *Fractional Dynamics on Networks and Lattices* (ISTE/Wiley, London, 2019).
20. Guo, Q., Cozzo, E., Zheng, Z. & Moreno, Y. Lévy random walks on multiplex networks. *Sci. Rep.* **6**, 37641 (2016).
21. Loaiza-Monsalve, D. & Riascos, A. P. Human mobility in bike-sharing systems: Structure of local and non-local dynamics. *PLoS ONE* **14**, e0213106 (2019).
22. de Nigris, S., Bautista, E., Abry, P., Avrachenkov, K. & Goncalves, P. Fractional graph-based semi-supervised learning. In *2017 25th European Signal Processing Conference (EUSIPCO)*, 356–360 (2017).
23. Zhao, Y., Weng, T. & Huang, D. Lévy walk in complex networks: An efficient way of mobility. *Physica A: Stat. Mech. Appl.* **396**, 212–223 (2014).
24. Weng, T., Small, M., Zhang, J. & Hui, P. Lévy walk navigation in complex networks: A distinct relation between optimal transport exponent and network dimension. *Sci. Rep.* **5**, 17309 (2015).
25. Weng, T. *et al*. Navigation by anomalous random walks on complex networks. *Sci. Rep.* **6**, 37547 (2016).
26. Estrada, E. *et al*. Random multi-hopper model: super-fast random walks on graphs. *J. Compl. Net.* **6**, 382–403 (2018).
27. Riascos, A. P., Michelitsch, T. M., Collet, B. A., Nowakowski, A. F. & Nicolleau, F. C. G. A. Random walks with long-range steps generated by functions of laplacian matrices. *Theory Exp.* **2018**, 043404 (2018).
28. Barthélemy, M. Spatial networks. *Phys. Rep.* **499**, 1–101 (2011).
29. Veloso, M., Phithakkitnukoon, S. & Bento, C. Urban mobility study using taxi traces. In *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*, TDMA **11**, 23–30 (ACM, New York, NY, USA, 2011).
30. Hoque, M. A., Hong, X. & Dixon, B. Analysis of mobility patterns for urban taxi cabs. In *2012 International Conference on Computing, Networking and Communications (ICNC)*, 756–760 (2012).
31. Wu, L., Zhi, Y., Sui, Z. & Liu, Y. Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PLoS One* **9**, e97010 (2014).
32. Santi, P. *et al*. Quantifying the benefits of vehicle pooling with shareability networks. *Proc. Natl. Acad. Sci. USA* **111**, 13290–13294 (2014).
33. Tachet, R. *et al*. Scaling law of urban ride sharing. *Sci. Rep.* **7**, 42868 (2017).
34. Vazifeh, M. M., Santi, P., Resta, G., Strogatz, S. H. & Ratti, C. Addressing the minimum fleet problem in on-demand urban mobility. *Nature* **557**, 534–538 (2018).

35. O'Keeffe, K. P., Anjomshoaa, A., Strogatz, S. H., Santi, P. & Ratti, C. Quantifying the sensing power of vehicle fleets. *PNAS* **116**, 12752–12757 (2019).
36. NYC borough boundaries, https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm.
37. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).
38. Klaus, A., Yu, S. & Plenz, D. Statistical analyses support power law distributions found in neuronal avalanches. *PLoS One* **6**, e19779 (2011).
39. Alstott, J., Bullmore, E. & Plenz, D. Powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS One* **9**, e85777 (2014).
40. powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions, https://pypi.org/project/powerlaw/.
41. Brin, S. & Page, L. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst* **30**, 107–117 (1998).
42. Gleich, D. Pagerank beyond the web. *SIAM Rev.* **57**, 321–363 (2015).
43. Ermann, L., Frahm, K. M. & Shepelyansky, D. L. Google matrix analysis of directed networks. *Rev. Mod. Phys.* **87**, 1261–1310 (2015).
44. Fortunato, S., Boguñá, M., Flammini, A. & Menczer, F. Approximating pagerank from in-degree. In Aiello, W., Broder, A., Janssen, J. & Milios, E. (eds.) *Algorithms and Models for the Web-Graph*, 59–71 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008).
45. Fortunato, S., Flammini, A., Menczer, F. & Vespignani, A. Topical interests and the mitigation of search engine bias. *PNAS* **103**, 12684–12689 (2006).
46. Litvak, N., Scheinhardt, W. R. W. & Volkovich, Y. Probabilistic relation between in-degree and pagerank. In Aiello, W., Broder, A., Janssen, J. & Milios, E. (eds.) *Algorithms and Models for the Web-Graph*, 72–83 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008).
47. Zhong, C., Arisona, S. M., Huang, X., Batty, M. & Schmitt, G. Detecting the dynamics of urban structure through spatial network analysis. *Int. J. Geogr. Inf. Sci.* **28**, 2178–2199 (2014).
48. Chin, W.-C.-B. & Wen, T.-H. Geographically modified pagerank algorithms: Identifying the spatial concentration of human movement in a geospatial network. *PLoS One* **10**, e0139509 (2015).
49. Huang, C.-Y., Chin, W.-C.-B., Wen, T.-H., Fu, Y.-H. & Tsai, Y.-S. Epirank: Modeling bidirectional disease spread in asymmetric commuting networks. *Sci. Rep.* **9**, 5415 (2019).
50. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P. & Tomkins, A. Geographic routing in social networks. *Proc. Natl. Acad. Sci. USA* **102**, 11623–11628 (2005).
51. Noulas, A., Scellato, S., Lambiotte, R., Pontil, M. & Mascolo, C. A tale of many cities: Universal patterns in human urban mobility. *PLoS One* **7**, e37027 (2012).
52. Pan, W., Ghoshal, G., Krumme, C., Cebrian, M. & Pentland, A. Urban characteristics attributable to density-driven tie formation. *Nat. Commun.* **4**, 1961 (2013).
53. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).
54. Yang, Y., Herrera, C., Eagle, N. & González, M. C. Limits of predictability in commuting flows in the absence of data for calibration. *Sci. Rep.* **4**, 5662 (2014).
55. Lenormand, M., Bassolas, A. & Ramasco, J. J. Systematic comparison of trip distribution laws and models. *J. Transp. Geogr.* **51**, 158–169 (2016).
56. NYC Data, https://opendata.cityofnewyork.us/.
57. Boeing, G. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Comput Environ Urban Syst* **65**, 126–139 (2017).
58. OSMnx Python Package, https://pypi.org/project/osmnx/.
59. Sagarra, O., Szell, M., Santi, P., Díaz-Guilera, A. & Ratti, C. Supersampling and network reconstruction of urban mobility. *PLoS One* **10**, e0134508 (2015).
60. OpenStreetMap contributors, https://www.openstreetmap.org.

Acknowledgements

This work was supported by PAPIIT-UNAM grant No. IN116220.

Author contributions

A.P.R. and J.L.M. designed the research, performed the research, and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.P.R. or J.L.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.