Uncovering the spatial structure of mobility networks

Thomas Louail[1,2], Maxime Lenormand[3], Miguel Picornell[4], Oliva García Cantú[4], Ricardo Herranz[4], Enrique Frias-Martinez[5], José J. Ramasco[3] & Marc Barthelemy[1,6]

The extraction of a clear and simple footprint of the structure of large, weighted and directed networks is a general problem that has relevance for many applications. An important example is seen in origin-destination matrices, which contain the complete information on commuting flows, but are difficult to analyze and compare. We propose here a versatile method, which extracts a coarse-grained signature of mobility networks, under the form of a 2×2 matrix that separates the flows into four categories. We apply this method to origin-destination matrices extracted from mobile phone data recorded in 31 Spanish cities. We show that these cities essentially differ by their proportion of two types of flows: integrated (between residential and employment hotspots) and random flows, whose importance increases with city size. Finally, the method allows the determination of categories of networks, and in the mobility case, the classification of cities according to their commuting structure.

[1] Institut de Physique Théorique, CEA-CNRS (URA 2306), Orme-des-Merisiers Batiment 774, F-91191 Paris, France. [2] Géographie-Cités, CNRS-Paris 1-Paris 7 (UMR 8504), 13 rue du four, FR-75006 Paris, France. [3] IFISC, Instituto de Física Interdisciplinar y Sistemas Complejos (CSIC-UIB), Campus Universitat de les Illes Balears, Palma de Mallorca E-07122, Spain. [4] Nommon Solutions and Technologies, calle Cañas 8, Madrid E-28043, Spain. [5] Telefonica Research, Madrid E-28050, Spain. [6] Centre d'Analyse et de Mathématique Sociales, EHESS-CNRS (UMR 8557), 190-198 avenue de France, FR-75013 Paris, France. Correspondence and requests for materials should be addressed to M.B. (email: marc.barthelemy@cea.fr).

The increasing availability of pervasive data in various fields has opened up exciting possibilities of renewed quantitative approaches to many phenomena. This is particularly true for cities and urban systems for which different devices at different scales produce a very large amount of data potentially useful to construct a 'new science of cities'[1].

A new problem we have to solve is to then extract useful information from these huge data sets. In particular, we are interested in extracting coarse-grained information and stylized facts that encode the essence of a phenomenon, and that any reasonable model should reproduce. Such meso-scale information helps us to understand the system, to compare different systems and also to propose models. This issue is particularly striking in the study of commuting in urban systems. In transportation research and urban planning, individuals' daily mobility is usually captured in origin-destination (OD) matrices, which contain the flows of individuals going from a point to another (see refs 2,3). An OD matrix thus encapsulates the complete information about the flow of individuals in a city, at a given spatial scale and for a specific purpose. It is a large network, and as such does not provide clear, synthetic and useful information about the structure of the mobility in the city. More generally, it is very difficult to extract high-level, synthetic information from large networks and methods such as community detection[4] and stochastic block modelling (see for example, refs 5–7) were recently proposed. Both these methods group nodes in clusters according to certain criteria, and nodes in a given cluster have similar properties (for example, in the stochastic block modelling nodes in a given group have similar neighbourhood). These methods are very interesting when one wants to extract meso-scale information from a network, but they are unable to construct expressive categories of links and propose a classification of weighted (directed) networks. This is particularly true in the case of commuting networks in cities, where edges represent flows of individuals that travel daily from their residential neighbourhood to their main activity area. Several types of links can be distinguished in these mobility networks; some constitute the backbone of the city by connecting major residential neighbourhoods to employment centres, while other flows converge from smaller residential areas to important employment centres, or diverge from major residential neighbourhoods to smaller activity areas. In addition, the spatial properties of these commuting flows are fundamental in cities and a relevant method should be able to take this aspect into account.

There is important literature in quantitative geography and transportation research that focuses on the morphological comparison of cities[8–11] and notably on multiple aspects of polycentrism, ranging from schematic pictures proposed by urban planners and architects[12] to quantitative case studies and contextualized comparisons of cities[13–15]. So far, most comparisons of large sets of cities have been based on morphological indicators[8,9] (built-up areas, residential density, number of sub-centres and so on) and aggregated mobility indicators[10,11] (motorization rate, average number of trips per day, energy consumption *per capita* per transport mode and so on), and have focused on the spatial organization of residences and employment centres. But these previous studies did not propose generic methods to take into account the spatial structure of commuting trips, which consist of both an origin and a destination. Such comparisons based on aggregated indicators thus fail to give an idea of the morphology of the city in terms of daily commuting flows. We still need some generic methods that are expressive in a urban context, and that could constitute the quantitative equivalent of the schematic pictures of city forms that have been pictured for so long by urban planners[12].

In this paper, we propose a simple and versatile method designed to compare the structure of large, weighted and directed networks. In the next section, we describe this method in detail. The guiding idea is that a simple and clear picture can be provided by considering the distribution of flows between different types of nodes. We then apply the method to commuting (journey to work) OD matrices of 31 cities extracted from a large mobile phone data set. We discuss the urban spatial patterns that our method reveals, and we compare these patterns observed in empirical data to those obtained with a reasonable null model that generates random commuting networks. Finally, the method allows determining categories of networks with respect to their structure, and here to classify cities according to their commuting structure. This classification highlights a clear relation between commuting structure and city size.

Results

Extracting coarse-grained information from OD matrices. For the sake of clarity, we will use here the language of OD matrices, but the method could easily be applied to any weighted and directed network from which we want to extract high-level information.

We assume that for a given city, we have the $n \times n$ matrix F_{ij} where n is the number of spatial units that compose the city at the spatial aggregation level considered (for example, a grid composed of square cells of size a, see Methods). This OD matrix F_{ij} represents the number of individuals living in location i and commuting to location j where they have their main, regular activity (work or school for most people). By convention, when computing the number of inhabitants and workers in each cell, we do not consider the diagonal of the OD matrix. This means that we omit the individuals who live and work in the same cell (considered as 'immobile' at this spatial scale).

To extract a simple signature of the OD matrix, we proceed in two steps. We first extract both the residential and the work locations with a large density—the so called 'hotspots' (see ref. 16). The number of residents of cell i is given by $\sum_{j \neq i} F_{ij}$ and its number of workers is given by $\sum_{j \neq i} F_{ji}$. The hotspots then correspond to local maxima of these quantities. It is important to note that the method is general, and does not depend on how we determine these hotspots.

Once we have determined the cells that are the residential and the work hotspots (some cells can possibly be both), we proceed to the second and main step of the method. We reorder the rows and columns of the OD matrix to separate hotspots from non-hotspots. We put the m residential hotspots on the top lines, and do the same for columns by putting the p work hotspots on the left columns. The OD matrix then becomes a four-quadrant matrix, where the flows F_{ij} are spatially positioned in the matrix with respect to their nature: on the top left the individuals that live in hotspots and work in hotspots; at the top right the individuals that live in hotspots and do not work in hotspots; at the bottom left individuals that do not live in hotspots but work in hotspots and finally in the bottom right corner the individuals that neither live or work in hotspots. For each quadrant we sum the number of commuters and normalize it by the total number of commuters in the OD matrix, which gives the proportion of individuals in each of the four categories of flows. In other words, for a given city, we reduce the OD matrix to a 2×2 matrix

$$\Lambda = \begin{pmatrix} I & D \\ C & R \end{pmatrix} \tag{1}$$

where

$$I = \sum_{i \in 1..m, j \in 1..p} F_{ij} \Big/ \sum_{i,j \in 1..n} F_{ij}$$

is the proportion of integrated flows that go from residential hotspots to work hotspots;

$$C = \sum_{i \in m+1..n, j \in 1..p} F_{ij} / \sum_{i,j \in 1..n} F_{ij}$$

is the proportion of convergent flows that go from random residential places to work hotspots;

$$D = \sum_{i \in 1..m, j \in p+1..n} F_{ij} / \sum_{i,j \in 1..n} F_{ij}$$

is the proportion of divergent flows that go from residential hotspots to random activity places;

$$R = \sum_{i \in m+1..n, j \in p+1..n} F_{ij} / \sum_{i,j \in 1..n} F_{ij}$$

is the proportion of random flows that occur 'at random' in the city, that is, that are going from and to places that are not hotspots.

By construction, we have $I, C, D, R \in [0,1]$ and $I + C + D + R = 1$. This matrix Λ is thus a very simple footprint of the OD matrix that gives an expressive picture of the structure of commuting in the city, as illustrated by Fig. 1.

Commuting data, hotspots and ICDR values. Large scale individual mobility networks are nowadays extracted from pervasive geolocated data, such as mobile phone, GPS, public transport cards or social apps data[17–21]. In particular, if an individual's mobile phone geolocated activity is available during a sufficiently long period of time, it is possible—under certain regularity conditions—to infer the most likely locations of her home and her workplace, and by aggregation to construct OD matrices[22–24]. Several parameters, however, impact the construction of OD matrices such as the nature of the data source (survey or user-generated geolocated data), or the spatial scale at which the OD matrix is built, which can be dictated by administrative units (divisions in wards, counties, municipalities and so on) or by technical reasons such as the density of antennas in the case of mobile phone data. Given this variety of data collection protocols,

Figure 1 | Illustration of the ICDR method. The method decomposes the commuting flows in the city in four categories: the integrated flows (I) from hotspot to hotspot, the convergent flows (C) to hotspots, the divergent flows (D) originating at hotspots and finally the random flows (R) which are neither starting nor ending at hotspots. For each city with its origin-destination matrix, we can compute the importance of each commuting flow category and get a simple picture of the mobility structure in the city.

it is thus particularly remarkable that when considering the commuting flows at a city scale, various sources of pervasive data provide a very similar mobility information when compared with the OD matrices built from surveys[24]. This result needs to be confirmed for other cities and countries, but it already opens the door to a systematic use of pervasive, geolocated data as a relevant substitute to traditional transport surveys.
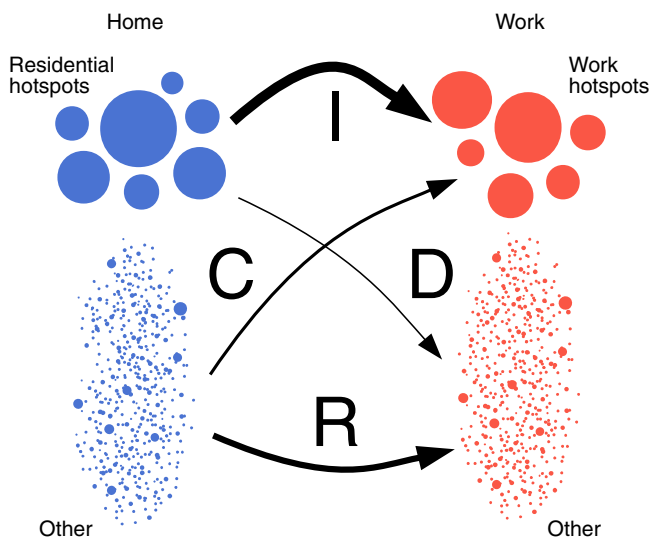
In the following, we apply our ICDR method to OD matrices that have been extracted from mobile phone records in 31 Spanish urban areas during a 5-week period (see the Methods section for details on the data set and the calculation of the OD matrices).

As described above, the first step consists in determining the hotspots. Several possible methods have been proposed in the literature[15,25,26], and we use here a parameter-free method based on the Lorenz curve of the densities that we have proposed in a recent study (see ref. 16 and the Methods section). Once we have determined both the origin (residential) hotspots and the destination (work) hotspots in each city, we first observe how their number scale with the population size of the city. Both these numbers for residential and employment hotspots scale sublinearly with the population size (see Supplementary Figs 4 and 5). The number of work hotspots grows significantly slower than the number of residential hotspots, showing that residential areas are (i) more dispersed in the city, and (ii) are more numerous than activity centres, as intuitively expected (see Supplementary Fig. 6 that displays the locations of home and work hotspots in four cities that exhibit different spatial organizations). We also note here that the sublinear scaling of the work hotspots confirms previous results obtained with a totally different data set (the number of employment centres in US cities)[27].
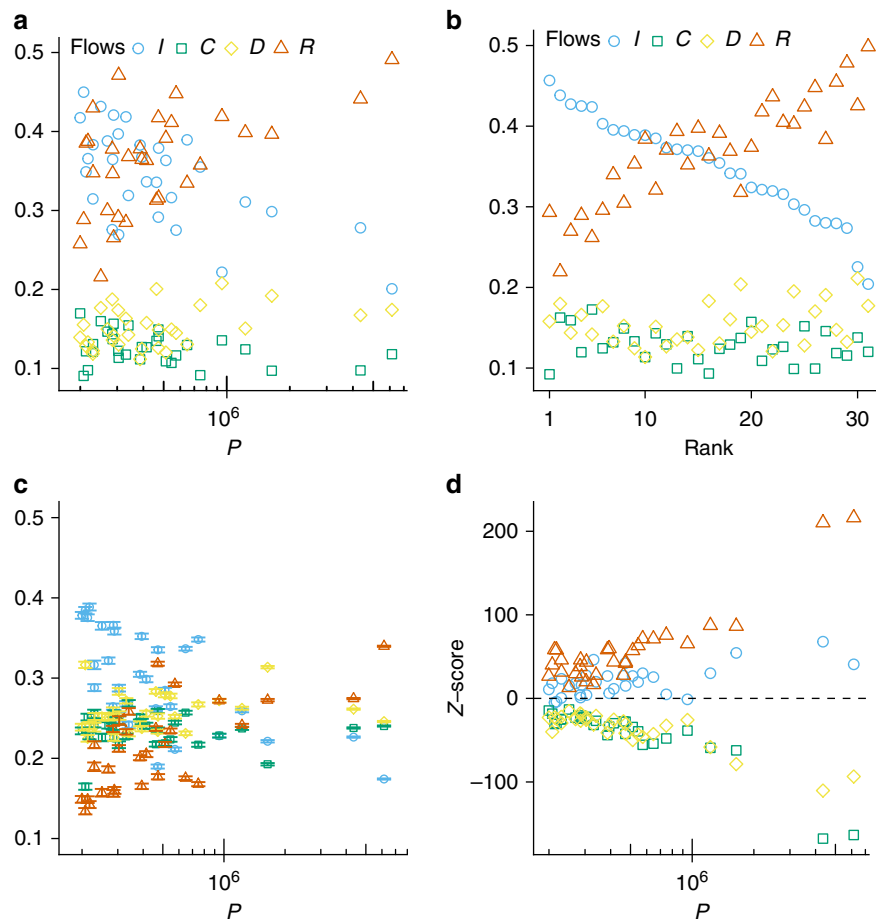
We now apply the second part of the method to calculate the I, C, D and R values for each OD matrix. For the 31 Spanish urban areas under study (see Supplementary Fig. 1), we obtain the values shown in Fig. 2. In Fig. 2a, we plot these values versus the population size of these cities. For this sample of cities, we see that globally the proportion I of individuals that commute from hotspot to hotspot decreases as the population size increases, while the proportion R of 'random' flows increases and the proportions C and D of convergent and divergent flows, respectively, seem surprisingly constant whatever the city size. In Fig. 2b we plot the same values but sorted by decreasing values of I, which shows clearly that the I and R values are the relevant parameters for distinguishing cities from each other.

We also notice that the values obtained for another spatial scale of data aggregation confirm this trend (see Supplementary Fig. 7). The decay of I flows ('integrated') flows in favour of R flows ('random') when P increases shows that the population growth among Spanish cities goes with a decentralization of both activity places and residences. As cities get bigger, their numbers of residential and employment hotspots grow (sublinearly), but these hotspots catch a smaller part of the commuting flows.

A null model. To evaluate to what extent the ICDR signatures of cities are characteristic of their commuting structure, we compare these values to the ones returned by a null model of commuting flows. For each city, we generate random OD matrices of the same size than the reference OD matrix but with random flows of individuals that preserve the in and out degree of each node (see Supplementary Note 5). Figure 2c shows the average values and s.d. obtained for 100 replications. In Fig. 2d, we plot the Z-scores of the I, C, D, and R values of each city when compared with the values $I^\star$, $C^\star$, $D^\star$ and $R^\star$ returned by our null model (for example, for the quantity R of the city i, the Z-score is given by $Z(R_i) = (R_i - \overline{R_i^\star})/\sigma(R_i^\star)$). Essentially, we observe that the

Figure 2 | Results for 31 Spanish cities. (**a**) *I* (integrated), *C* (convergent), *D* (divergent) and *R* (random) values versus population size for 31 Spanish urban areas. (**b**) Same ICDR values as in **a** but sorted by decreasing order of *I* (note that by definition, we have for each city $I + C + D + R = 1$). It is remarkable that *I* and *R* dominate and seem almost sufficient to distinguish cities, while *C* and *D* are almost constant whatever the city size (see Supplementary Fig. 7 for the values obtained with another size a of grid cells). (**c**) *I*, *C*, *D* and *R* average values and s.d. obtained for 100 replications of a null model, where the inflow and outflow at each node are kept constant while flows are randomly distributed at random between nodes. (**d**) *Z*-scores obtained by comparing the empirical data and the values returned by the null model. Large values of *Z*-scores show that the actual commuting networks cannot be considered as resulting from connecting the nodes at random. The *I*, *C*, *D* and *R* values of a specific city are then a signature of its structure.

Z-scores of *I* and *R* are positive and large, while those of *C* and *D* are negative (and large in absolute value). Also as cities grow, the *Z*-scores of *I* and *R* increase while those of *C* and *D* decrease. These results demonstrate that the larger a city, and the less random it appears. This is in contrast with the naive expectation that the larger a city, the more disordered is the structure of individuals' mobility. For large cities, there seems to be a commuting backbone, which cannot result from purely random movements of individuals. This backbone is the footprint of the city's structure and history, and probably results from strong constraints and efficiency considerations.

Robustness. Since our method first requires to determine origin and destination hotspots, one could argue that the interpretation of the *I*, *C*, *D* and *R* values will crucially depend on the particular method chosen to define these hotspots. The identification of hotspots is a problem that has been broadly discussed in urban economics (see Supplementary Note 3). Roughly speaking, starting from a spatial distribution of densities, the goal is to identify the local maxima and amounts to choose a threshold $\rho^\star$ for the density ρ of individuals: a cell i is a hotspot if the local density of people is such that $\rho(i) > \rho^\star$. To test the impact of the choice of a particular threshold $\rho^\star$ on the resulting ICDR values, we measure the sensitivity of these values to the density threshold

$\rho^\star$ (see Supplementary Fig. 8). As expected, the lower the density threshold, the larger the number of origin and destination hotspots, and consequently the larger *I* and the smaller *R*. In contrast, changing the density threshold has little impact on the *C* and *D* terms. More importantly, the conclusions drawn from the comparison of the ICDR values across cities remain the same; whatever the density threshold $\rho^\star$ chosen to define residential/employment hotspots, we still observe the same qualitative behaviour: a decay of integrated flows (*I*) in favour of 'random' flows (*R*), when the population size increases.

Distance of each type of flows. We now want to characterize spatially these different flows and the relation between city size and the commuting distances travelled by individuals. In each city, we compute the average distance travelled by individuals per type of flows *I*, *C*, *D* and *R*. The resulting average distances measured in data are plotted in red on Fig. 3a. We observe that the average distance for all categories of flows increases with population size, an expected effect as the city's area also grows with population size. We also observe that the average distance of convergent flows *C* increases faster than for other types of flows (we note that the average distance associated to convergent flows increases more than the distance associated to divergent flows *D*, showing that these flows are not symmetric as one could have
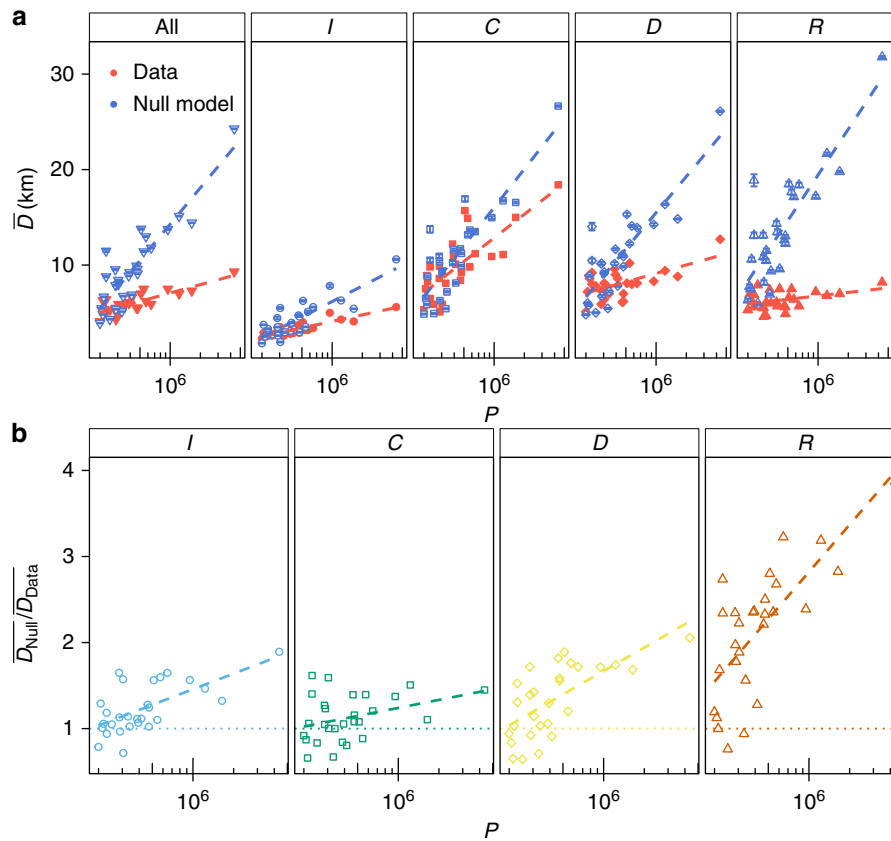
Figure 3 | Distance per type of flow and population size. (**a**) Average commuting distance versus population size, per type of flows; (**b**) ratio $\overline{D_{\text{Null}}}/\overline{D_{\text{Data}}}$ per type of flows. The LouBar criteria[16] is used here to define residential and employment hotspots.
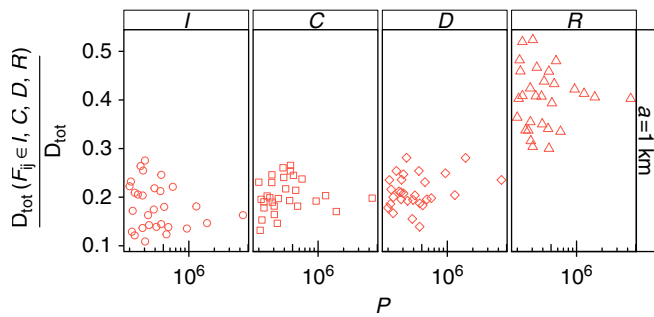
Figure 4 | Total commuting distance. Contributions to the total commuting distance of each type of flows versus population size. We observe that the variations are small and that the largest contribution comes from the *R* flows.

naively expected). This result means that for this set of Spanish cities, commuters from small residential areas to important activity centres travel on average a longer distance than all other individuals. This observation could be an indication that for our set of cities, residential areas have expanded while activity centres remained at their location, leading to longer commuting distances.

Another interesting information is provided by the comparison of distances measured in the data with average distances measured from the random OD matrices generated by the null model. The average distances associated to the null model are plotted in blue in Fig. 3a. We see that for all types of flows the distances measured in the empirical data are shorter than those generated by the null model. This is another clear indication of

the spatial organization of individual flows in cities. It also highlights the importance of the travel time budget in the residential locations choice. Remarkably enough, the distance of convergent flows (*C*) is both the largest and the one that increases the fastest with population, indicating a low degree of efficiency.

The comparison of this behaviour with the null model leads to interesting results. In Fig. 3b, we plot the ratio $\overline{D_{\text{Null}}}/\overline{D_{\text{Data}}}$ for the four types of flow. Values <1 indicate that the average commuting distance generated by the null model is shorter than the distance observed in the city. Surprisingly, we observe that small cities display a value <1, indicating the lesser importance of space at this short scale. We also see that this ratio increases faster for random flows (*R*) than for the others (*D*, *C*, *I*), suggesting a remarkable spatial structure of these *R* flows.

We also consider the fraction of total commuting distance by type of flow (Fig. 4). We see that for each type of flows, their respective fraction is constant and independent of city size. With the LouBar hotspots detection method[16] (see Supplementary Fig. 3) and with a grid of 1 km² cells, we measure that roughly 40% of the total commuting distance is made on random flows while the other types represent each about 20%. This result shows that the method is able to identify where most of the commuting distance is travelled. In particular, we see that the natural, obvious flows (*I*) from residential centres to activity centres are not the most important ones, and that the decentralization of commuting flows seems to be the rule for the Spanish cities in our sample (see Supplementary Figs 10 and 11 for various aggregation scales).

Classification of cities. Finally, the ICDR signature of their OD matrix allows to cluster cities with respect to the structure of their commuting patterns. We measure the euclidian distance between

the cities' ICDR signatures and we then perform a hierarchical cluster analysis. Figure 5 shows the dendrogram resulting from the classification. Four well-separated clusters are identified on this dendrogram, and Table 1 gives the average value of each term along with the average population of the cities composing the cluster. Remarkably these summary statistics show that largest cities are clustered together and are characterized by a larger proportion of 'random' flows (R) of individuals both living and working in parts of the city that are not the dominant residential and activity centres. This can be interpreted as an increased facility in bigger urban areas to commute from any part of the city to any other part. Further studies on other cities and countries are needed at this stage to discuss the relevance of the proposed classification.

It is also important to test the robustness of this classification and we show that introducing a reasonable amount of noise in the OD matrices does not change the classification (see Methods and Supplementary Fig. 12). This sensitivity test confirms that the clustering is robust against possible errors in the data source and in the extraction of the mobility networks. The classification of cities based on their *ICDR* values is also robust to a change of the method used to define residential and employment hotspots (see Supplementary Fig. 13).
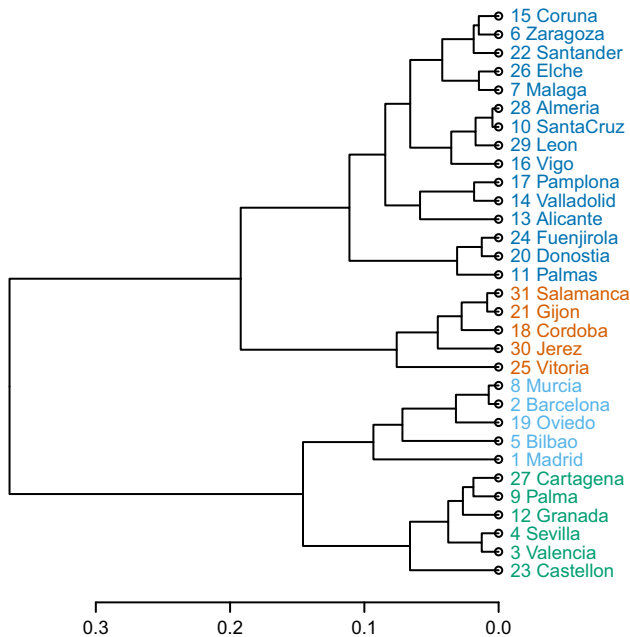
Figure 5 | Classification of cities. Dendrogram resulting from the hierarchical clustering on cities based on their ICDR values. In front of each city name, we indicate its rank in the hierarchy of population sizes. The largest cities are clustered together. As cities get bigger, the 'random' component (R) of their commuting flows increases, which signals that it is easier to commute from any place to any other in large cities.

Discussion

We have proposed a method to extract high-level information from large, weighted and directed networks, such as OD matrices. This method relies on the identification of origin and destination hotspots, and this first step can be performed with any reasonable criterion. The important second step consists of aggregating flows of four different types, depending whether they start and end from/to a hotspot or not.

We have applied this method to commuting networks extracted from mobile phone data available in 31 Spanish cities. The method has allowed us to highlight several remarkable patterns in the data:

- Independently of the density threshold chosen to determine hotspots, the proportion of integrated flows (I) decreases with city size, while the proportion of random flows (R) increases;
- On average and for all cities considered here, individuals who live in residential main hubs and who work in employment main hubs (I flows) travel shorter distances than the others (C, D, R flows);
- When the city size increases, the largest impact is on convergent flows (C) of individuals living in smaller residential areas (typically in the suburbs) and commuting to important employment centres;
- The classification of cities based on the ICDR values leads to groups with consistent population size, highlighting a clear relationship between the population size of cities and their commuting structure.

In addition, the comparison with a null model led to interesting conclusions. Flows in cities display a high level of spatial organization and as the population size of the city grows, the increase of the Z-scores of I, C, D, R shows that the structure of the mobility is more and more specific, and far from a random organization. We note that an interesting direction for future research would be to find some analytical arguments using simple models of city organization for estimating these flows, and how they vary with population size.

Our coarse-graining method provides a large scale picture of individual flows in the city. In this respect it could be particularly useful for validating synthetic results of urban mobility models (such as ref. 28), and for comparing different models. An accurate modelling of mobility is indeed crucial in a large number of applications, including the important case of epidemic spreading, which needs to be better understood, in particular at the intra-urban level[29,30].

It would also be interesting to apply the proposed method to other mobility data sets, with different time and spatial scales, to test the robustness of these commuting patterns. Another direction for future studies could be to inspect the time evolution of the I, C, D and R values for data sets describing travel to work journeys over several decades (such as national travel surveys). The method could also be applied at larger spatial scales, for

Table 1 | Classification of cities.

Cluster	Cities	$\bar{P}$	$\bar{I}$	$\bar{R}$	$\bar{D}$	$\bar{C}$
Orange	Salamanca, Gijon, Cordoba,…	255,330	0.43	0.27	0.16	0.14
Dark blue	La Coruña, Zaragoza, Santander, Elche,…	392,970	0.37	0.36	0.15	0.13
Green	Cartagena, Palma, Granada,…	732,992	0.31	0.41	0.16	0.13
Light blue	Murcia, Barcelona, Bilbao, Madrid,…	2,463,551	0.25	0.46	0.17	0.12

Average ICDR values and average population sizes of the cities composing each of the four clusters represented in Fig. 5. As the population grows the proportion of random flows increases while the proportion of integrated decreases. The weights of convergent and divergent flows stay constant among the groups.

example, to capture dominant effects in international migration flows. More generally, we believe that an important feature of the ICDR method is its versatility, as it could be applied to any type of data that is naturally represented by a weighted and directed network.

Methods

Data. The data set used for our analysis comprises 55 days of aggregated and anonymized records for 31 urban areas of more than 200,000 inhabitants. No individual information or records were available for this study. The records included the set of base transceiver stations (communication antennas, BTS) used for the communications as presented in the call detail records (CDR). A CDR is produced for each active phone event, including call/sms sending or reception. The number of anonymized users represents on average 2% of the total population and at most 5% of the total population. These percentages are almost the same for all the urban areas. From the CDR data obtained for 20 weekdays (from mondays to thursdays only), we extracted home and work places for all the anonymized mobile phone users in the data set. The output of this processing phase is an OD commuting matrix for each urban area, at the scale of the BTS point pattern. To facilitate the calculations and the comparison of the results between different cities, the OD matrices are then transposed on regular square-cell grids of varying size a (see Supplementary Note 2).

Extraction of OD matrices from mobile phone data. In order to extract OD matrices from phone calls, we select a subset of users with a mobility displaying a sufficient level of statistical regularity. For this analysis, we considered commuting patterns during workdays only. The users' home and work locations are identified as the Voronoi cells, which are the most frequently visited on weekdays by each user between 2000 and 0700 h (home) and between 0900 and 1700 h (work). We assume that there must be a daily travel between the home and work locations of each individual. Users with a call activity >40% of the days under study at home or work are considered as valid. We then aggregate the complete flow of users and construct the OD matrix with the flows between a Voronoi cell classified as home and another cell classified as a workplace. Since the Voronoi areas do not exactly match the grid cells (see for example Supplementary Fig. 2), we use a transition matrix to change the spatial scale of the OD matrix, that is to transform the F_{ij} values of the OD matrix where i and j are Voronoi cells into $F'_{i'j'}$ values where i' and j' are the cells of a regular grid (see Supplementary Note 2).

Spatial scale of the OD matrix. The OD matrix is the standard object in mobility studies and transport planning[2], and contains information about movement of individuals in a given area. More precisely, an OD matrix is a $n \times m$ matrix, where n is the number of different 'origin' zones, m is the number of 'destination' zones and F_{ij} is the number of people commuting from place i to place j during a given period of time. In transport surveys, the size of the OD matrix depends on the spatial scale at which the mobility data has been collected. Traditionally, the zones that are used to partition the city are the administrative units, whose size can vary from census and electoral units to whole departments or states, depending on the purpose for building the OD matrix.

In this study, we applied our ICDR method to cities divided in square cells that are smaller than administrative units, allowing for a better spatial resolution. In the case of OD matrices extracted from CDR mobile phone data, the maximal resolution corresponds to the BTS (antennas) point pattern. The ICDR method proposed in this paper does, however, depend on a particular spatial scale and can be applied on OD matrices available at coarser spatial resolutions as well. For a given territory, the results obtained with the ICDR method—the I, C, D and R values in the first place—will obviously depend on the spatial resolution and will also depend on the method used to define hotspots. It is important to note that when the ICDR method is used for comparing cities, the spatial resolution and the hotspots identification method should be the same for all cities (see Supplementary Notes 4 and 6 and Supplementary Figs 8 and 9 for results obtained with another hotspots delimitation method, and see Supplementary Fig. 7 for results obtained when considering another spatial scale of aggregation).

Robustness of the classification of cities. To ensure that the classification of cities based on their ICDR matrices is robust, we introduce a noise in the flows F_{ij} (for all the 31 cities). We focus on the case where the workplace of an individual can be modified and where the number of individuals living in each cell i is kept constant. More precisely, the noise is introduced as follows:

1. We pick up a uniform random positive integer g, the number of individuals whose workplace is reshuffled. This number g varies from 1 to $N = \sum F_{ij}$ the total number of commuters in the city.
2. We repeat g times the following operation: we pick up randomly a residence and a workplace (a couple of values (i, j)) and move one individual from her workplace j to put another randomly chosen workplace j': $F_{ij} \rightarrow F_{ij} - 1$ and $F_{ij'} \rightarrow F_{ij'} + 1$

The parameter of this workplace reshuffling is then $f = g/N$. To evaluate how much the classification of cities is affected by this noise, we compute the Jaccard index J_I between the reference classification of cities in k groups and the classification in k groups obtained for the noisy OD matrices. The Jaccard index measures the similarity of two partitions P and P' of same size:

$$J_I = \frac{a}{a + b + c} \quad (2)$$

where

- a: number of city pairs that are in the same group for both P and P'
- b: number of city pairs that are in different groups in P but in the same one in P'
- c: number of city pairs that are in the same group in P but not in P' (or conversely)

The Jaccard index J_I is in the range [0, 1] and the closest it is to 1, the larger the similarity between P and P'.

We generate 100 noisy matrices for each value of f and compute the average value $\bar{J}_I$ of the Jaccard index. This average value encodes the distance between the reference partition P of cities in k groups and the partitions of cities in k groups obtained for the noisy OD matrices. Supplementary Fig. 12 shows the values of $\bar{J}_I$ versus the proportion f of reshuffled individuals, for different number of groups k. The red shaded rectangle on each panel corresponds to the mean value ± s.d. obtained for 1,000 replications of a null model, in which permutations of cities among the k groups are randomly performed. We observe here that for up to 20% of reshuffled individuals, the average value $\bar{J}_I$ obtained is always significantly larger than the one obtained for the null model, indicating that the classification is robust even for important values of noise.

References

1. Batty, M. *The New Science of Cities* (MIT Press, 2013).
2. Ortuzar, J. D. & Willumsen, L. G. Modelling transport 4th edn (Wiley, 2011).
3. Weiner, E. *Urban Transportation Planning in the United States* (US Department of Transportation, 1986).
4. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486,** 75–174 (2010).
5. Faust, F. & Wasserman, S. *Social Network Analysis* (Cambridge Univ. Press, 1992).
6. Karrer, B. & Newman, M. E. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83,** 016107 (2011).
7. Aicher, C., Jacobs, A. Z. & Clauset, A. Learning Latent Block Structure in Weighted Networks. Preprint at http://arXiv.org/abs/1404.0431 (2014).
8. Tsai, Y. H. Quantifying urban form: compactness versus sprawl. *Urban. Stud.* **42,** 141–161 (2005).
9. Guérois, M. & Pumain, D. Built-up encroachment and the urban field: a comparison of forty european cities. *Env. Plan. A* **40,** 2186–2203 (2008).
10. Schwarz, N. Urban form revisited - selecting indicators for characterising european cities. *Landscape Urban Plan.* **96,** 29–47 (2010).
11. Le Néchet, F. Urban spatial structure, daily mobility and energy consumption: a study of 34 european cities. *Cybergeo* 580 (2012).
12. Bertaud, A. & Malpezzi, S. The Spatial Distributio of Population in 48 World Cities: implications for Economies in Transition (World Bank Report, 2003).
13. Salomon, I., Bovy, P. & Orfeuil, J. P. *A Billion Trips a Day: Tradition and Transition in European Travel Patterns* (Kluwer Academic Publishers, 1993).
14. Cattan, N. *Cities and Networks in Europe: a Critical Approach of Polycentrism* (John Libbey Eurotext, 2007).
15. Berroir, S., Mathian, H., Saint-Julien, T. & Sanders, L. in *Modelling Urban Dynamics* (eds Desrosiers, F. & Thériault, M.) 1–25 (ISTE-Wiley, 2011).
16. Louail, T. *et al.* From mobile phone data to the spatial structure of cities. *Sci. Rep.* **4,** 5276 (2014).
17. Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z. & González, M. C. Unravelling daily human mobility motifs. *J. R. Soc. Interface* **10,** 20130246 (2013).
18. Roth, C., Kang, S. M., Batty, M. & Barthelemy, M. Structure of urban movements: polycentric activity and entangled hierarchical flows. *PLoS ONE* **6,** e15923 (2011).
19. 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).
20. 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).
21. 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).
22. Kung, K. S., Sobolevsky, S. & Ratti, C. Exploring universal patterns in human home/work commuting from mobile phone data. *PLoS ONE* **9,** e96180 (2014).
23. Jiang, S. *et al.* in Proceedings of the 2nd ACM SIGKDD (International Workshop on Urban Computing, 2013).

24. Lenormand, M. *et al.* Cross-checking different sources of mobility information. *PLoS ONE* **9,** e105184 (2014).
25. McMillen, D. P. Nonparametric employment subcenter identification. *J. Urban Econ.* **50,** 448–473 (2001).
26. Griffith, D. A. Modelling urban population density in a multi-centered city. *J. Urban Econ.* **9,** 298–310 (1981).
27. Louf, R. & Barthelemy, M. Modeling the polycentric transition of cities. *Phys. Rev. Lett.* **111,** 198702 (2013).
28. Eubank, S. *et al.* Modelling disease outbreaks in realistic urban social networks. *Nature* **429,** 180–184 (2004).
29. Balcan, D. *et al.* Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl Acad. Sci. USA* **106,** 21484–21489 (2009).
30. Dalziel, B. D., Pourbohloul, B. & Ellner, S. P. Human mobility patterns predict divergent epidemic dynamics among cities. *Proc. R. Soc. B* **280,** 20130763 (2013).

Acknowledgements

Author contributions

T.L. designed the study, processed and analyzed the data and wrote the manuscript; M.L. processed and analyzed the data; O.G.C. and M.P. processed the data; R.H. and J.J.R. coordinated the study; E.F.-M. obtained and processed the data; M.B. coordinated and designed the study, and wrote the manuscript. All authors read, commented and approved the final version of the manuscript.

Additional information