INTERFACE

rsif.royalsocietypublishing.org

Research

Cite this article: Leo Y, Fleury E, Alvarez-Hamelin JI, Sarraute C, Karsai M. 2016 Socioeconomic correlations and stratification in social-communication networks. *J. R. Soc. Interface* **13**: 20160598.
http://dx.doi.org/10.1098/rsif.2016.0598

Received: 29 July 2016
Accepted: 17 November 2016

Subject Category:
Life Sciences–Physics interface

Subject Areas:
environmental science

Keywords:
socioeconomic status, social networks, stratification, rich clubs, human mobility

Author for correspondence:
Márton Karsai
e-mail: marton.karsai@ens-lyon.fr

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3584051.

THE ROYAL SOCIETY
PUBLISHING

Socioeconomic correlations and stratification in social-communication networks

Yannick Leo[1], Eric Fleury[1], J. Ignacio Alvarez-Hamelin[2,3], Carlos Sarraute[4] and Márton Karsai[1]

[1]Univ Lyon, ENS de Lyon, INRIA, CNRS, UMR 5668, IXXI, 69364 Lyon, France
[2]Univerisidad de Buenos Aires, Facultad de Ingeniería, Av. Paseo Colón 850, C1063ACV, Buenos Aires, Argentina
[3]CONCIET, Buenos Aires, Argentina
[4]Grandata Labs, Bartolome Cruz 1818 Vicente Lopez, Buenos Aires, Argentina

JIA-H, 0000-0002-2910-9320; MK, 0000-0001-5382-8950

The uneven distribution of wealth and individual economic capacities are among the main forces, which shape modern societies and arguably bias the emerging social structures. However, the study of correlations between the social network and economic status of individuals is difficult due to the lack of large-scale multimodal data disclosing both the social ties and economic indicators of the same population. Here, we close this gap through the analysis of coupled datasets recording the mobile phone communications and bank transaction history of one million anonymized individuals living in a Latin American country. We show that wealth and debt are unevenly distributed among people in agreement with the Pareto principle; the observed social structure is strongly stratified, with people being better connected to others of their own socioeconomic class rather than to others of different classes; the social network appears to have assortative socioeconomic correlations and tightly connected 'rich clubs'; and that individuals from the same class live closer to each other but commute further if they are wealthier. These results are based on a representative, society-large population, and empirically demonstrate some long-lasting hypotheses on socioeconomic correlations, which potentially lay behind social segregation, and induce differences in human mobility.

1. Introduction

Socioeconomic imbalances, which universally characterize all modern societies [1,2], are partially induced by the uneven distribution of economic power between individuals. Such disparities are among the key forces behind the emergence of social inequalities [2,3], which in turn leads to social stratification and spatial segregation in social structures characterized by correlations between the social network, living environment and socioeconomic status of people. Although this hypothesis was drawn a long time ago [4], the empirical observation of spatial, socioeconomic and structural correlations in large social systems has been difficult as it requires simultaneous access to multimodal characters for a large number of individuals. Our aim in this study is to find evidence of social stratification through the analysis of a combined large-scale anonymized dataset that discloses simultaneously the social interactions, frequent locations and the economic status of millions of individuals.

The identification of socioeconomic classes is among the historical questions in the social sciences with several competing hypothesis proposed on their structure and dynamics [5]. One broadly accepted definition identifies lower, middle and upper classes [6–10] based on the socioeconomic status of individuals. These classes can be further used to indicate correlations characterizing the social system. People who live in the same neighbourhood may belong to the same class, and may have similar levels of education, jobs, income, ethnic background, and may even share common political views. These similarities together with

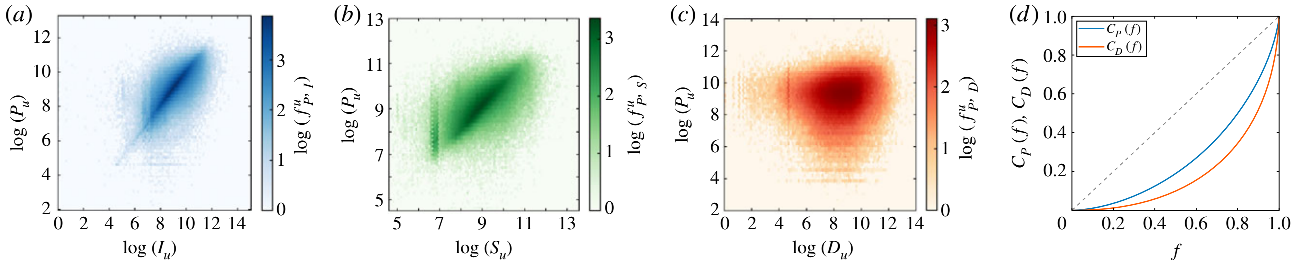

Figure 1. Correlations and distributions of individual economic indicators. The heat maps show correlations between the AMP P_u and (*a*) average income I_u, (*b*) average salary S_u and (*c*) average monthly debt D_u for (*a*) 625 412 (*b*) 389 567 and (*c*) 339 288 customers who have (accordingly) both corresponding measures available. Colours in panels (*a*–*c*) depict the logarithm of the fraction of customers with the given measures. (*d*) Cumulative distributions of P_u (blue line) and D_u (orange line) as functions of sorted fraction f of individuals. Distributions were measured for 6 002 192 (resp.339 288) individuals from whom AMP (resp. AMD) values were available. Dashed line shows the case of perfectly balanced distribution.

homophily, i.e. the tendency of people to build social ties with similar others [11,12], strongly influence the structure of social interactions and also have indisputable consequences on the global social network. The coexistence of social classes and homophily may lead to a strongly stratified social structure where people of the same social class tend to be better connected among each other, while connections between different classes are less frequent than one would expect from structural characteristics only [4,13,14]. These correlations may further determine the living environment and mobility of people leading to spatial segregation and specific commuting patterns characterizing people from similar social classes [15–17].

The observation of such correlations should be possible through the analysis of the social structure [18]. Research on social networks has recently been accelerated through the advent of new technologies which allow the collection of detailed digital footprints of interactions of large numbers of people [19,20]. These advancements have allowed us to observe that social networks appear with heterogeneous connection patterns, are structurally, spatially and temporally correlated [21,22], and to identify various social mechanisms driving their evolution [23,24]. However, although such datasets may contain some information about individual characteristics, they commonly miss one important dimension: they do not provide any direct estimator of the economic status of people, which could strongly influence their connection preferences and may determine the social position of an individual in the global social network. Coarse-grained information details about people's economic status are typically provided as statistical census measures without disclosing the underlying social structure, or by social surveys [25] covering a small and less representative population.

In this paper, we aim to close this gap through the analysis of a combined dataset collecting the social interactions, proxy location and economic situation of a large set of individuals. More precisely, we analyse the transaction and purchase history coupled with time-resolved, spatially detailed mobile phone interactions of millions of anonymized inhabitants of a Latin American country over eight months (for a detailed data description, see Data and material). After introducing precise indicators of economic status, we show that not only individual income but also debt is distributed unevenly in accordance with the Pareto principle. Through the detection of homophilic correlations in the social structure, we provide strong empirical evidence of the stratified intra- and inter-class structure of the social network, and the existence of assortative socioeconomic correlations and 'rich clubs'. Finally, we present quantitative results about the relative spatial distribution and typical commuting distances of people from different socioeconomic classes.

2. Results

The full description of one's socioeconomic status is rather difficult as it is characterized not only by quantitative features but also related to one's social or cultural capital [26], reputation or professional skills. However, we can estimate socioeconomic status by assuming a correlation between one's social position and economic status, which can be approximated by following the network position and financial development of people. This approach in turn not only gives us a measure of an individual's socioeconomic status but can also help us to draw conclusions about the overall distribution of socioeconomic potential in the larger society.

2.1. Economic status indicators

Our estimation of an individual's economic status is based on the measurement of consumption power. We use a dataset which contains the amount and type of daily debit/credit card purchases, monthly loan amounts and some personal attributes such as age, gender and zip code of billing address of approximately six million anonymized customers of a bank in the studied country over eight months (for further details see Data and material). In addition, for a smaller subset of clients, the data provide the precise salary and total monthly income that we use for verification purposes as explained later.

By following the purchase history of each individual, we estimate their economic position from their average amount of debit card purchases. More precisely, for an individual u who spent a total amount of $P_u(t)$ in month t, we estimate his/her average monthly purchase (AMP) as

$$P_u = \frac{\sum_{t \in T} P_u(t)}{|T|_u}, \qquad (2.1)$$

where $|T|_u$ corresponds to the number of active months of the user (with at least one purchase). In order to verify this individual economic indicator, we check its correlations with other indicators, such as the salary S_u (defined as the average monthly salary of individual u over the observation period T) and the income I_u (defined as the average total monthly income including salary and other incoming bank transfers). We find strong correlations between individual AMP P_u and income I_u with a Pearson correlation coefficient $r \approx 0.758$ ($p < 0.001$, s.e. $= 7.33 \times 10^{-4}$) (for correlation heat map, see figure 1*a*), and also between P_u and salary S_u with $r \approx 0.691$ ($p < 0.001$, s.e. $= 9.695 \times 10^{-4}$) (figure 1*b*). Note that direct economic indicators, such as I_u and S_u, are available only for a smaller subset of users (for exact numbers, see figure 1), thus for this study we decided to use P_u because this measure is available for the whole set of users.

At the same time, we are interested in an equivalent indicator, which estimates the financial commitments of individuals. We define the average monthly debt (AMD) of an individual u by measuring

$$D_u = \frac{\sum_{t \in T} d_u(t)}{|T|_u}, \tag{2.2}$$

where $d_u(t)$ indicates the debt of individual u in month $t \in T$ and $|T|_u$ is the number of active months where the user had debt. Arguably, individual debt could depend on the average income and thus on the AMP of a person due to the loaning policy of the bank. Interestingly, as demonstrated in figure 1*c*, we found weak correlations between AMP and AMD with a small coefficient $r \approx 0.104$ ($p < 0.001$, s.e. $= 2.48 \times 10^{-3}$), which suggests that it is worth studying these two indicators independently.

2.2. Overall socioeconomic imbalances

The distribution of an individual economic indicator may disclose signs of socioeconomic imbalances at the population level. This hypothesis was first suggested by V. Pareto and later became widely known as the law named after him [27]. The present data provide a straightforward way to verify this hypothesis through the distribution of individual AMP. We measured the normalized cumulative function of AMP for f fraction of people sorted by P_u in an increasing order

$$C_P(f) = \frac{1}{\sum_u P_u} \sum_f P_u. \tag{2.3}$$

We computed this distribution for the 6 002 192 individuals assigned with AMP values. This function shows (figure 1*d* blue line) that AMP is distributed with a large variance, i.e. indicating large economical imbalances just as suggested by Pareto's law. A conventional way to quantify the variation of this distribution is provided by the Gini coefficient G [28], which characterizes the deviation of the $C_P(f)$ function from a perfectly balanced situation, where wealth is evenly distributed among all individuals (diagonal dashed line in figure 1*d*). In our case, we found $G_P \approx 0.461$, which is relatively close to the World Bank reported value $G = 0.481$ for the studied country [29], and corresponds to a Pareto index [30] of $\alpha = 1.315$. This observation indicates a 0.73 : 0.27 ratio characterizing the uneven distribution of wealth, i.e. 27% of people are responsible for 73% of the total monthly purchases in the observed population. Note that these values are close to the values $G = 0.6$ and 80 : 20, which were suggested by Pareto.

At the same time, we have characterized the distribution of individual AMD by measuring the corresponding $C_D(f)$ function as shown in figure 1*d* (orange line) for 339 288 individuals for whom AMD values were available. It indicates even larger imbalances in the case of debt with a Gini coefficient $G_D \approx 0.627$ and $\alpha = 1.140$ indicating 19% of the population to be actually responsible for 81% of the overall debt in the country. This observation suggests that Pareto's hypothesis holds not only for the distribution of purchases but also for debt. Note that a similar distribution of debt of bankrupt companies has been reported [31].

2.3. Class definition and demographic characters

The economic capacity of individuals arguably correlates with their professional occupation, education level and housing, which in turn determine their social status and environment. At the same time, status homophily [11,12], i.e. people's tendency to associate with others of similar social status, has been argued to be an important mechanism that drives the creation of social ties. Our hypothesis is that these two effects, diverse socioeconomic status and status homophily, potentially lead to the emergence of a stratified structure in the social network where people of the same social class tend to be better connected among themselves than with people from other classes. A similar hypothesis had been suggested earlier [32] but its empirical verification had been impossible until now as this would require detailed knowledge about the social structure and precise estimators of individual economic status. In the following, our main contribution is to clearly identify signatures of social stratification in a representative society-level dataset, which contains information on both the social network structure and the economic status of people.

In order to investigate signatures of social stratification, we combine the bank transaction data with data disclosing the social connections between the bank's customers. To identify social ties, we use a mobile communication dataset, provided by one mobile phone operator in the country, with a customer set that partially overlaps with the user set found in the bank data (for details on data matching policy, see Data and material). To best estimate the social network, we connect people who communicated with each other at least once via calling or SMS during the observation period of 21 months between January 2014 and September 2015, but we remove non-human actors, such as call centres and commercial communicators by using a recursive filtering method. For the purpose of our study, we select all mobile phone users who appear as customers in the bank dataset and take the largest connected component of the intersection graph. After this procedure, we obtain a social network with $|E| = 1\,960\,239$ links and $N = 992\,538$ nodes, each corresponding to an individual with a valid non-zero AMP value P_u. For further details about the datasets, their combinations, filtering and network construction, see Data and material.

Taking each individual in the selected social network, we assign each of them to one of $n = 9$ socioeconomic classes based on their individual AMP values. This classification is defined by sorting individuals by their AMP, taking the cumulative function $C_P(f)$ of AMP and cutting it into n segments such that the sum of AMP in each class is equal to $(\sum_u P_u)/n$ (as shown in figure 2*a*). Our selection of nine distinct classes is based on the common three-stratum model [6,7], which identifies three main social classes (lower, middle and upper), and three sub-classes for each of them [14]. More importantly, this way of classification relies merely on individual economic estimators, P_u, and naturally partitions individuals into classes with decreasing sizes, and increasing $\langle P \rangle$ *per capita* average AMP values for richer groups (for exact values, see figure 2*b*). (To assign purchase values in USD, we used the daily average currency rate (17.90 MXN/USD) on 2 March 2016.) To explore the demographic structure of the classes, we used data on the age and gender of customers.

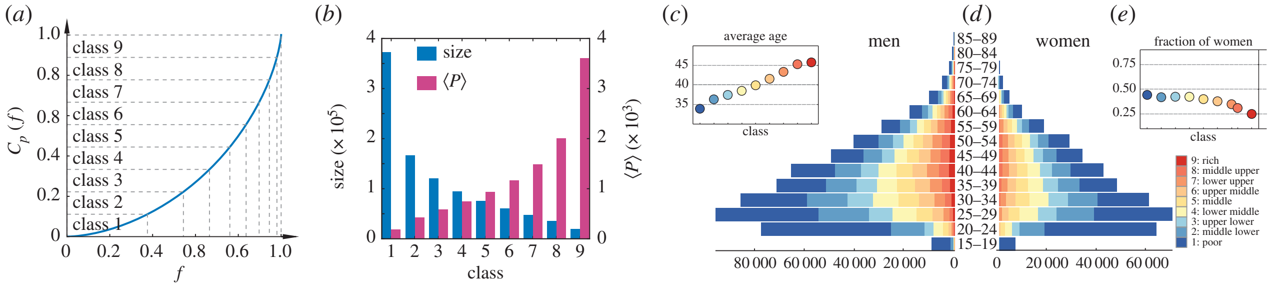

Figure 2. Social class characteristics (*a*) schematic demonstration of user partitions into nine socioeconomic classes by using the cumulative AMP function $C_P(f)$. Fraction of individuals belonging to a given class (*x*-axis) has the same sum of AMP $\left(\sum_u P_u\right)/n$ (*y* axis) for each class. (b) Number of individuals (blue), and the average AMP $\langle P\rangle$ (to assign purchase values in USD we used the daily average currency rate (17.90 MXN/USD) on the 2nd March 2016) per individual (pink) in different classes. (*c*) Average age of different classes. (*d*) Age pyramids for men and women with colours indicating the corresponding socioeconomic groups and with bars proportional to absolute numbers. (*e*) Fraction of women in different classes.

We have drawn the population pyramids for men and women in figure 2*d* with coloured bars indicating the number of people in a given social class at a given age. We found a positive correlation between social class and average age, suggesting that people in higher classes are also older on average (figure 2*c*). In addition, our data verify the presence of gender imbalance as the fraction of women varies from 0.45 to 0.25 going from lower to upper socioeconomic classes (figure 2*e*).

2.4. Structural correlations and social stratification

Using the above-defined socioeconomic classes and the social network structure, we turn to look for correlations in the inter-connected class structure. To highlight structural correlations, such as the probability of connectedness, we use a randomized reference system. It is defined as the corresponding configuration network model structure where we take the original social network, select random pairs of links and swap them without allowing multiple links and self-loops. Hence, the degree of each of the four nodes involved in the swap remains unchanged. In order to remove any residual correlations, we repeated this procedure 5 × $|E|$ times. This degree-preserving randomization keeps the number of links, individual degrees (and hence any degree–wealth correlations), individual economic indicators P_u, and the assigned class of people unchanged, but destroys any higher-order structural correlations in the social structure and consequently also between socioeconomic layers. In each case, we repeat this procedure 100 times and present results averaged over the independent random realizations. Taking the original (resp. randomized) network, we count the number of links $|E(s_i,\ s_j)|$ (resp. $|E_{rn}(s_i,\ s_j)|$) connecting people in different classes s_i and s_j. After repeating this procedure for each pair of classes in both networks, we take the fraction

$$L(s_i, s_j) = \frac{|E(s_i, s_j)|}{|E_{rn}(s_i, s_j)|}, \tag{2.4}$$

which gives us how many times more (or less) links are present between classes in the original structure when compared with the randomized one. Note that in the randomized structure the probability that two people from given classes are connected depends only on the number of social ties of the individuals and the size of the corresponding classes, but is independent of the effect of potential structural correlations. This way the comparison of the original and random structures highlights structural patterns induced by anything other than node degrees. Such patterns could emerge due to status homophily, degree–degree correlations (as we study later here and in the electronic supplementary material), or due to triadic closure, communities, motifs and any other structural correlations that one could think of.

From the chord diagram visualization of this measure in figure 3*a*, we can draw several conclusions. Note that for better visual presentation in figure 3*a*, we have normalized $L(s_i,\ s_j)$ and thus chord width indicates relative values $\widetilde{L}_{s_i}(s_j) = L(s_i, s_j)/\sum_{s_j} L(s_i, s_j)$ when compared with the origin class s_j (as also explained in the figure caption). First, after sorting the chords of a given class s_i in a decreasing $L(s_i,\ s_j)$ order, chords connecting a class to itself (self-links) always appear at top (or top second) positions of the ranks. At the same time, other top positions are always occupied by chords connecting to neighbouring social classes. These two observations (better visible in figure 3*a* insets) indicate strong effects of status homophily and the existence of stratified social structure where people from a given class are the most connected with similar others from their own or from neighbouring classes, while connections with individuals from remote classes are least frequent. A second conclusion can be drawn by looking at the sorting of links in the middle and lower upper classes ($S4$–$S8$). As demonstrated in the inset of figure 3*a*, people prefer to connect upward and tend to hold social ties with others from higher social classes rather than with people from lower classes.

These conclusions can be further verified by looking at other representations of the same measure. First, we show a heat map matrix representation of equation (2.4) (figure 3*b*), where $L(s_i,\ s_j)$ values are shown with logarithmic colour scales. This matrix has a strong diagonal component verifying that people of a given class are always better connected among themselves (red) and with others from neighbouring groups, while social ties with people from remote classes are largely under-represented (blue) when compared with the expected value provided by the random reference model. This again indicates the presence of homophily and the stratified structure of the socioeconomic network. The upward-biased inter-class connectivity can also be concluded here from the increase of the red area around the diagonal by going towards richer classes. These conclusions are even more straightforward from figure 3*c* where the $L(s_i,\ s_j)$ is shown for three selected classes (1, poor; 5, middle and 9, rich). These curves clearly indicate the connection preferences of the selected classes. Moreover, they show that the richest people appear with the strongest homophilic preferences as their class is approximately 2.25 times better connected among each other than

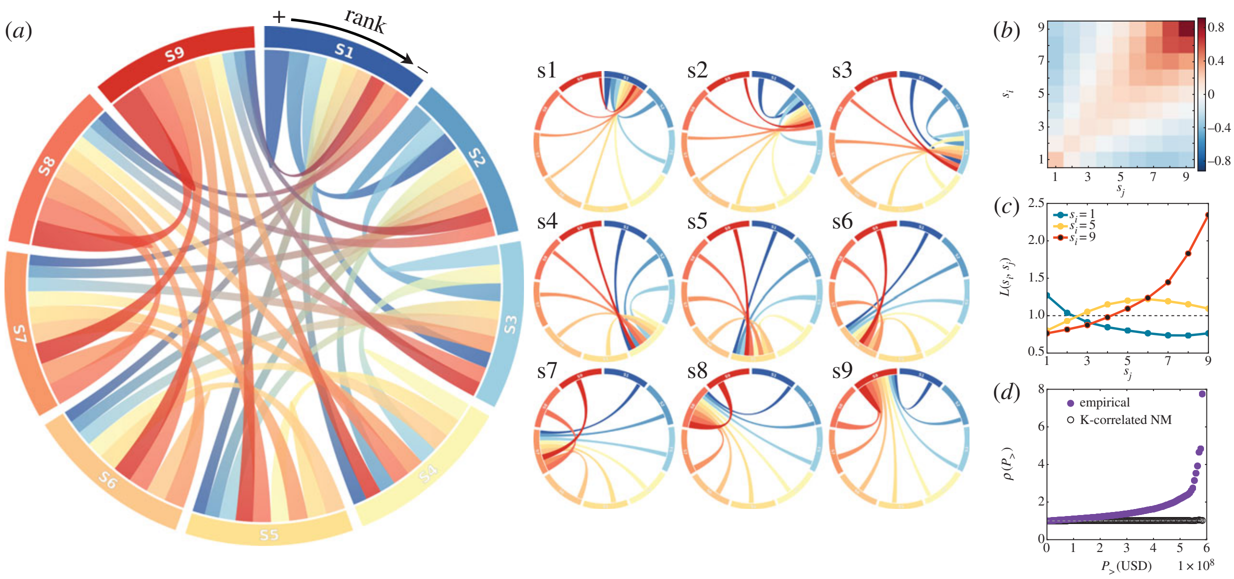

Figure 3. Structural correlations in the socioeconomic network. (*a*) Chord diagram of connectedness of socioeconomic classes s_i, where each segment represents a social class s_i connected by chords with width proportional to the corresponding inter-class link fraction $\tilde{L}_{s_i}(s_j)$, and using gradient colours matched with opposite ends s_j. Note that the $\tilde{L}_{s_i}(s_j) = L(s_i, s_j)/\Sigma_{s_j} L(s_i, s_j)$ normalized fraction of $L(s_i, s_j)L(s_i, s_j)$ (in equation (2.4)) was introduced here to assign equal segments for each class for better visualization. Chords for each class are sorted in decreasing width order in the direction shown above the main panel. On the minor chord diagrams of panel (*a*), graphs corresponding to each class are shown with non-gradient link colours matching the opposite end other than the selected class. (*b*) Matrix representation of $L(s_i, s_j)$ (for definition see equation (2.4)) with logarithmic colour scale. (*c*) The $L(s_i, s_j)$ function extracted for three selected classes (1 (blue), 5 (yellow) and 9 (red)). Panels (*a*–*d*) provide quantitative evidence on the stratified structure of the social network and the upward-biased connections of middle classes. (*d*) 'Rich-club' coefficient $\rho(P_>)$ (definition see equation (2.6)) based on the empirical (purple), and a degree-correlated null model (black) networks. On the individual level, the richest people of the population appear to be eight times more densely connected than expected randomly.

expected by chance, on the expense of weaker connectivity to remote classes. This effect is somewhat weaker for middle classes, which function as bridges between poor and rich classes, but apparently upward-biased towards richer classes. This set of results directly verifies our earlier conjectures that the structure of the socioeconomic network is strongly stratified and builds up from social ties, whose creation is potentially driven by status homophily, and determined by the socioeconomic characteristics of individuals.

However, one can argue that the observed stratified structure can be simply the consequence of simultaneously present degree–degree and degree–wealth correlations. More precisely, if the degree of an individual is highly correlated with its economic status and at the same time the network is strongly assortative (i.e. people prefer to connect to other people with similar degrees), we may observe similar effects as in figure 3*a*–*c*. To rule out this possibility, we completed an extensive correlation analysis, which showed us that no strong effects of degree–degree correlations can be detected and that the degree and wealth of individuals are very weakly correlated. To further clarify the effects of these correlations, we defined another null model similar to the configuration network model, but where degree–degree correlations were preserved. In this model, instead of selecting link pairs randomly for swapping, we select a link and one of its ends randomly, and choose another link randomly where the degree of one of the ending nodes is equal to the degree of the selected end of the first link. Swapping the other ends of the links (with potentially different degrees) will result in two other links between nodes of the original degrees but connected randomly otherwise (for further details, see the electronic supplementary material). Using this null model, we demonstrated that simultaneously present degree–degree and degree–wealth correlations cannot explain the observed stratified structure (for results, see the electronic supplementary material).

The above observations further suggest that the social structure may show assortative correlations in terms of socioeconomic status at the individual level. In other words, richer people may be better connected among themselves than one would expect them by chance and this way they form tightly connected 'rich clubs' in a structure similar to the suggestion of Mills [33]. This can be verified by measuring the rich-club coefficient [34,35], after we adjust its definition to our system as follows. We take the original social network structure, sort individuals by their AMP value P_u and remove them in an increasing order from the network (together with their connected links). At the same time, we keep track of the density of the remaining network defined as

$$\phi(P_>) = \frac{2L_{P_>}}{N_{P_>}(N_{P_>} - 1)}, \qquad (2.5)$$

where $L_{P_>}$ and $N_{P_>}$ are the number of links and nodes remaining in the network after removing nodes with P_u smaller than a given value $P_>$. In our case, we consider $P_>$ as a cumulative quantity going from 0 to $\sum_u P_u$ with values determined just as in the case of $C_P(f)$ in figure 2*a* but now using 100 segments. At the same time, we randomize the structure using a configuration network model and by removing nodes in the same order, we calculate an equivalent measure $\phi_{\text{rn}}(P_>)$ as defined in equation (2.5) but in the uncorrelated structure. For each randomization process, we used the same parameters as earlier and calculated the average density $\langle\phi_{\text{rn}}\rangle(P_>)$ of the networks over 100 independent realizations. Using the two density functions, we define the 'rich-club' coefficient as

$$\rho(P_>) = \frac{\phi(P_>)}{\langle\phi_{\text{rn}}\rangle(P_>)}, \qquad (2.6)$$

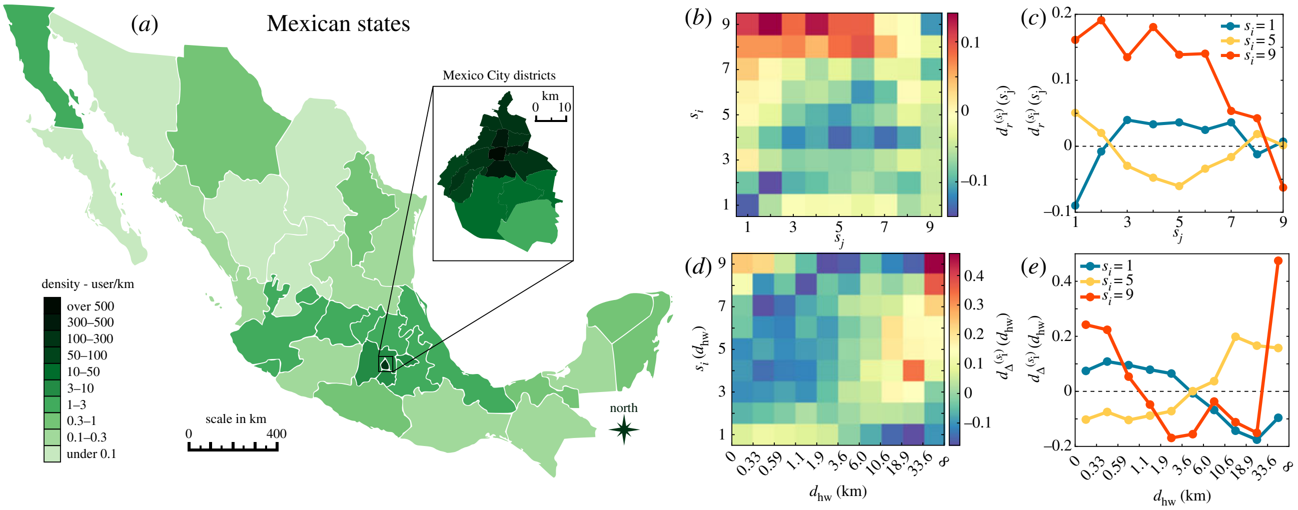

Figure 4. Spatial socioeconomic correlations. (*a*) State level population distribution of individuals based on their zip locations. Inset depicts a zoom on the capital district. Information details depicted here were entirely obtained from the used dataset. The map representation was generated by using an open source code available at https://gist.github.com/diegovalle/5843688github.com/diegovalle (no copyright reserved) and shape files openly available at http://www.inegi.org.mxwww.inegi.org.mx (no copyright reserved). (*b*) Relative average geodesic distances for different classes using the measure $d_r^{s_i}(s_j)$ defined in equation (2.8). (*c*) The same $d_r^{s_i}(s_j)$ functions as on panel (*b*) shown for a selected set of classes (1-poor (blue), 5-middle (yellow), 9-rich (red)). (*d*) $d_\Delta^{s_i}(d_{\mathrm{hw}})$ differences between commuting distance distributions calculated for different classes and for the whole population. *x*-scale depicts in logarithmic values of d_{hw} commuting distances. (*e*) The same $d_\Delta^{s_i}(d_{\mathrm{hw}})$ functions as on panel (*d*) shown for a selected set of classes (1-poor (blue), 5-middle (yellow), 9-rich (red)).

which indicates how many times the remaining network of richer people is denser connected than expected from the reference model. In our case (figure 3*d* purple symbols), the rich-club coefficient increases monotonously with $P_>$ and grows rapidly once only the richer people remain in the network. At its maximum, it shows that the richest people are approximately eight times more connected in the original structure than in the uncorrelated case. This provides direct evidence about the existence of tightly connected 'rich clubs' [33], and the presence of strong assortative correlations in the social structure on the level of individuals in terms of their socioeconomic status. Note that this measure also suggests that the observed 'rich clubs' were not induced by degree–wealth correlations. The connectedness of nodes in the randomized structure were actually determined merely by their degrees, and because we kept wealth–degree correlations, the wealth-sorted removal process shows exactly the expected density of remaining richer nodes assuming only their original degree but no other correlations. This way the fraction of the two network density curves, i.e. the rich-club coefficient, actually characterizes exactly the effect of status homophily when compared with the randomized case where only degrees and degree–wealth correlations determined the connectedness of the network.

In addition, to rule out the possibility that our observation was induced by positive degree–degree correlations, we performed another randomization of the network, where we kept node degrees, degree–degree and degree–wealth correlations but removed any other structural correlations. This randomization procedure preserving degree–degree correlations is identical to the one we defined earlier and in the electronic supplementary material. To measure the corresponding rich-club coefficient function, we substituted in the numerator of equation (2.6) the residual network density function measured in this new degree-correlated null model using the same wealth-sorted removal sequence as earlier. Results in figure 3*d* (black symbols) show that the obtained rich-club coefficient appears approximately as a constant function around one. This way it demonstrates that the entangled effects of degree–degree and degree–wealth correlations cannot explain the emergence of 'rich clubs' observed in the empirical case. The network, which conserves degrees and these two correlations, emerges with a structure just as the network, which conserves degrees and degree–wealth correlations only. Consequently, the observed increasing rich-club coefficient in the case of the empirical structure is induced by status homophily or other tie creation mechanisms and not by degree–degree or degree–wealth correlations.

2.5. Spatial correlations between socioeconomic classes

As we discussed earlier, the economic capacity of an individual strongly determines the possible places he/she can afford to live, arguably leading to somewhat homogeneous neighbourhoods, districts, towns and regions occupied by people from similar socioeconomic classes. This effect may translate to correlations in the spatial distribution of socioeconomic classes in relation with each other. To study such correlations, we use three different types of geographical information extracted for individuals from the data: the zip code of the reported billing address; the home; and work locations estimated from call activity logs (for details, see Data and material). To give an overall image about the spatial distribution of the investigated users, we use their zip location and assign them in different states of the country as shown in figure 4*a*. Importantly, the observed population distribution correlates well with census data [36] with coefficient $r = 0.861$ $(p < 0.001)$ on the state level, which indicates that our data record a fairly unbiased sample of the population in terms of distribution in space.

To quantify spatio-socioeconomic correlations, we measure the relative average geodesic distance between classes. More precisely, we take all connected individuals $(u, v) \in E$ belonging to classes $u \in s_i$ and $v \in s_j$, respectively and measure the geodesic distance $d_{\mathrm{geo}}^{\mathrm{zip}}(a, b)$ between their zip locations. Using these values, we calculate the average geodesic distance between any pairs of socioeconomic classes as

$$\langle d_{\mathrm{geo}}(s_i, s_j)\rangle = \frac{1}{|E(s_i, s_j)|} \sum_{\substack{(u,v)\in E \\ u\in s_i, v\in s_j}} d_{\mathrm{geo}}^{\mathrm{zip}}(u, v), \tag{2.7}$$

where $|E(s_i, s_j)|$ assigns the number of links between nodes in classes s_i and s_j. Note that because the social network is undirected the measure defined in equation (2.7) is symmetric, i.e. $\langle d_{\mathrm{geo}}(s_i, s_j)\rangle = \langle d_{\mathrm{geo}}(s_j, s_i)\rangle$. Subsequently, we calculate the average distance between nodes from class s_i and any of their neighbours $\langle d_{\mathrm{geo}}(s_i)\rangle$ to derive

$$d_r^{s_i}(s_j) = \frac{\langle d_{\mathrm{geo}}(s_i, s_j)\rangle - \langle d_{\mathrm{geo}}(s_i)\rangle}{\langle d_{\mathrm{geo}}(s_i)\rangle}. \tag{2.8}$$

This measure is not symmetric anymore and gives us the relative average geodesic distance between individuals in s_i to individuals in other classes s_j when compared with the average distance of individuals s_i from any of their connected peers. Results are presented as a heat map matrix in figure 4*b* where the diagonal component suggests a peculiar correlation. It shows that the relative average distance is always minimal (and negative) between individuals of the same class s_i. This means that people tend to live relatively the closest to similar others from their own socioeconomic class as to individuals from different classes, independently in which class they belong to. This is even more visible in figure 4*c* after extracting the $d_r^{s_i}(s_j)$ curves (corresponding to rows in figure 4*b*) for three selected classes. It highlights that while people of the poorest class live relatively the closest to each other, rich people tend to leave relatively the furthest from anyone from lower socioeconomic classes. These correlations are very similar to ones we already observed in the social structure suggesting that the stratified structure and spatial segregation may have similar roots. They are determined by the entangled effects of economic status and status homophily, together with other factors such as ethnicity or other environmental effects, which we cannot consider here.

The socioeconomic status of people may also correlate with their typical commuting distances (between home and work), a question that has been studied thoroughly during the last few decades. Some of these studies suggest a positive correlation between economical status (income) and the distance people travel every day between their home and work locations [37–39]. Such correlations were partially explained by the positive payoff between commuting farther for better jobs, while keeping better housing conditions. On the other hand, recent studies suggest that such trends may change nowadays as in central metropolitan areas, where the better job opportunities are concentrated, became more expensive to live and thus occupied by people from richer classes [40,41]. Without going into detail, we looked for overall signs of such correlations by using the estimated home (ℓ_{h}) and work (ℓ_{w}) locations of individuals from different classes. For each individual, we measure a commuting distances as $d_{\mathrm{hw}} = |\ell_{\mathrm{h}} - \ell_{\mathrm{w}}|$ and compute the $P_{s_i}(d_{\mathrm{hw}})$ distributions for everyone in a given s_i class, together with the $P_{\mathrm{all}}\ (d_{\mathrm{hw}})$ distribution considering all individuals. For each class, we are interested in

$$d_{\Delta}^{s_i}(d_{\mathrm{hw}}) = P_{s_i}(d_{\mathrm{hw}}) - P_{\mathrm{all}}(d_{\mathrm{hw}}), \tag{2.9}$$

i.e. the difference between the corresponding distributions at each distance d_{hw}. This measure is positive (resp. negative) if more (resp. less) people commute at a distance d_{hw} when compared with the overall distribution, thus indicating whether people of a given class are over (under)represented at a given distance. Interestingly, our data are in agreement with both of the above-mentioned hypotheses, as seen in figure 4*d* where we show $d_{\Delta}^{s_i}(d_{\mathrm{hw}})$ for each class as a heat map. There, poorer people are over represented in shorter distances while this trend is shifted towards larger distances (see right skewed yellow component in figure 4*d*) as going up in the class hierarchy. This continues until we reach the richest classes (8 and 9) where the distance function becomes bimodal assigning that more people of these classes tend to live very far or very close to their work places when compared with expectations considering the whole population. This is even more visible in figure 4*e* where selected $d_{\Delta}^{s_i}(d_{\mathrm{hw}})$ functions are depicted for selected classes.

3. Discussion

In this paper, we have investigated socioeconomic correlations through the analysis of a coupled dataset of mobile phone communication records and bank transaction history for millions of individuals over eight months. After mapping the social structure and estimating individual economic capacities, we addressed four different aspects of their correlations: (i) we showed that individual economic indicators such as AMPs and also debts are unevenly distributed in the population in agreement with the Pareto principle; (ii) after grouping people into nine socioeconomic classes, we detected effects of status homophily and showed that the socioeconomic network is stratified as people most frequently maintain social ties with people from their own or neighbouring social classes; (iii) we observed that the social structure is upward-biased towards wealthier classes and show that assortative correlations give rise to strongly connected 'rich clubs' in the network; (iv) finally, we demonstrated that people of the same socioeconomic class tend to live closer to each other when compared with people from other classes, and found a positive correlation between their economic capacities and the typical distance they use to commute.

Even though our study is built on large and detailed data, the used data cover only partially the population of the investigated country. However, as we demonstrated above, for population-level measures, such as the Gini coefficient and spatial distribution, we obtained values close to independently reported cases, and thus our observations may generalize in this sense. In addition, the question remains how well mobile phone call networks approximate real social structure. A recent study [42] demonstrated that real social ties can be effectively mapped from mobile call interactions with precision up to 95%. However, it is important to keep in mind that the poorest social class of the society is probably underrepresented in the data as they may have no access to bank services and/or do not hold mobile phones. Datasets simultaneously disclosing the social structure and the socioeconomic indicators of a large number of individuals are still very rare. However, several promising directions have been proposed lately to estimate socioeconomic status from communication behaviour on regional level [43–45] or even for individuals [46], just to mention a few. In future works, these methods could be used to generalize our results to other countries using mobile communication datasets. Here, our aim was to report some general observations in this direction using directly estimated individual economic indicators. Our overall motivation was to empirically verify some long-standing hypotheses and to explore a common ground between hypothesis-driven and data-driven research addressing social phenomena.

4. Data and material

4.1. Mobile communication data

Communication data used in our study record the temporal sequence of 7 945 240 548 call and SMS interactions of 111 719 360 anonymized mobile phone users for 21 months (between January 2014 and September 2015) in Mexico. Each call detailed record contains the time, unique caller and callee IDs, the direction and duration of the interaction, and the cell tower location of the client(s) involved in the interaction. Other mobile phone users, who are not clients of the actual provider also appear in the dataset with unique IDs. All unique IDs are anonymized as explained below, thus individual identification of any person is impossible from the data. Using this dataset, we constructed a large social network where nodes were users (whether clients or not of the actual provider), while links were drawn between them if they interacted (via call or SMS) at least once during the observation period. In order to filter out call services and other non-human actors from the social network, after construction we recursively removed all nodes (and connected links) who appeared with either in-degree $k_{\mathrm{in}} = 0$ or out-degree $k_{\mathrm{out}} = 0$. We repeated this procedure recursively until we received a network where each user had $k_{\mathrm{in}}, k_{\mathrm{out}} > 0$, i.e. made at least one outgoing and received at least one incoming communication events during the nearly 2 years of observation. After construction and filtering the network remained with 82 453 814 users connected by 1 002 833 289 links, which were considered to be undirected after this point.

4.2. Credit and purchase data

To estimate individual economic indicators, we used a dataset provided by a single bank in the studied country. These data record financial details of 6 002 192 of people assigned with unique anonymized identifiers over eight months from November 2014 to June 2015. The data provide time varying customer variables as the amount and type of their daily debit/credit card purchases, their monthly loan measures, and static user attributes as their billing postal code (zip code), their age, and gender. In addition, for a subset of clients we have the records of monthly salary (38.9% of users) and income (62.5% of users) defined as the sum of their salaries and any incoming bank transactions. Note that the observation period of the bank credit information falls within the observation period of the mobile communication dataset, this way ensuring the largest possible overlap between the sets of bank and mobile phone customers.

4.3. Location data

We used two types of location data for a set of customers. We used the zip code of billing address of bank customers (also called zip location). We also estimated the work and home locations for a set of users using geo-localized mobile communication events. To determine home (resp. work) locations, we looked for the most frequented locations during nights and weekends (resp. during daylight at working days). From the total 992 538 individuals, we found 990 173 with correct zip codes, and 94 355 with detectable home and work locations (with at least 10 appearances at each location). Each method has some advantages and disadvantages. While frequency-dependent locations are more precise, they strongly depend on the activity and regularity of users in terms of mobility. On the other hand, zip codes provide a more coarse-grained information about the location of individuals but they are assumed to be more reliable due to reporting constraints to the bank and because they do not depend on the call activity of individuals.

4.4. Combined datasets and security policies

A subset of IDs of the anonymized bank and mobile phone customers were matched. The matching, data hashing and anonymization procedure were carried out through direct communication between the two providers (bank and mobile provider) and were approved by the national banking commission of the country. This procedure was done without the involvement of the scientific partners. After this procedure only anonymized hashed IDs were shared disallowing the direct identification of individuals in any of the datasets. Owing to the signed non-disclosure agreements and the sensitive nature of the datasets, it is impossible to share them publicly.

This way of combining of the datasets allowed us to simultaneously observe the social structure and estimated economic status of the connected individuals. The combined dataset contained 999 456 IDs, which appeared in both corpuses. However, for the purpose of our study we considered only the largest connected component of this graph containing IDs valid in both data corpuses. This way we operate with a connected social graph of 992 538 people connected by 1 960 242 links, for all of them with communication events and detailed bank records available.

Authors' contributions. All authors participated in the design of the project and the writing of the manuscript. Y.L., M.K. and E.F. designed the measures; Y.L. and M.K. performed the data analysis. All authors reviewed the manuscript.

Competing interests. We have no competing interests.

Funding. We acknowledge the support from the SticAmSud UCOOL project, INRIA,and the CODDDE (ANR-13-CORD-0017-01) and SoSweet (ANR-15-CE38-0011-01) ANR projects.

Acknowledgements. We thank M. Fixman and J. Brea for assistance with the dataset, J. Saramäki and J. P. Chevrot for useful discussions and the anonymous reviewers for their constructive comments.

References

1. Piketti T. 2014 *Capital in the twenty-first century*. Cambridge, MA: Harvard University Press.
2. Sernau S. 2013 *Social inequality in a global age*. Beverley Hills, CA: SAGE Publications.
3. Hurst CE. 2015 *Social inequality*, 8th edn. London, UK: Pearson Education.
4. Grusky DB. 2011 Theories of stratification and inequality. In *The concise encyclopedia of sociology* (eds G Ritzer, JM Ryan) pp. 622–624. Oxford, UK: Wiley-Blackwell.
5. Giddens A, Ociepka F, Zujewicz W. 1973 *The class structure of the advanced societies*. London, UK: Hutchinson.
6. Akhbar-Williams T. 2010 Class structure. In *Encyclopedia of African American popular*

culture, vol. 1 (ed. JC Smith), pp. 320–323. Westport, CT: Greenwood.

7. Brown DF. 2009 Social class and status. In *Concise encyclopedia of pragmatics* (ed. JL Mey), 953. Amsterdam, The Netherlands: Elsevier.
8. Stark R. 2007 *Sociology*. Belmont, CA: Thompson/Wadsworth.
9. Gilbert D. 2002 *The American class structure: in an age of growing inequality*. Beverley Hills, CA: Pine Forge Press.
10. Stiglitz J. 2012 *The price of inequality*. New York, NY: Norton.
11. McPherson M, Smith-Lovin L, Cook JM. 2001 Birds of a feather: homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444. (doi:10.1146/annurev.soc.27.1.415)
12. Lazarsfeld PF, Merton RK. 1954 Friendship as a social process: a substantive and methodological analysis. In *Freedom and control in modern society*, vol. 18 (eds M Berger, T Abel, CH Page), pp. 18–66. Van Nostrand series in sociology. New York, NY: Van Nostrand.
13. Doob CB. 2016 *Social inequality and social stratification in US society*. London, UK: Routledge.
14. Saunders P. 1990 *Social class and stratification*. London, UK: Routledge.
15. Carra G, Mulalic I, Fosgerau M, Barthelemy M. 2016 Modelling the relation between income and commuting distance. *J. R. Soc. Interface* **13**, 119. (doi:10.1098/rsif.2015.0315)
16. Sim A, Yaliraki SN, Barahona M, Stumpf MPH. 2015 Great cities look small. *J. R. Soc. Interface* **12**, 20150315. (doi:10.1098/rsif.2016.0306)
17. Iceland J, Wilkes R. 2006 Does socioeconomic status matter? Race, class, and residential segregation. *Soc. Probl.* **53**, 248–273. (doi:10.1525/sp.2006.53.2.248)
18. Wasserman S, Faust K. 1994 *Social network analysis: methods and applications*. Cambridge, UK: Cambridge University Press.
19. Lohr S. 2012 *The age of big data*. New York, NY: New York Times.
20. Lazer D *et al*. 2009 Computational social science. *Science* **323**, 721–723. (doi:10.1126/science.1167742)
21. Abraham A, Hassanien AE, Smasel V. 2010 *Computational social network analysis: trends, tools and research advances*. New York, NY: Springer.
22. Newman MEJ. 2003 The structure and function of complex networks. *SIAM Rev.* **45**, 167–256. (doi:10.1137/S003614450342480)
23. Hedström P, Swedberg R. 1998 *Social mechanisms, an analytical approach to social theory*. Cambridge, UK: Cambridge University Press.
24. Holme P, Liljeros F. 2015 Mechanistic models in computational social science. *Front. Phys*. **3**, 78. (doi:10.3389/fphy.2015.00078)
25. Campbell KE, Marsden PV, Hurlbert JS. 1986 Social resources and socioeconomic status. *Soc. Netw.* **8**, 97–117. (doi:10.1016/S0378-8733(86)80017-X)
26. Bourdieu P. 1984 *Distinction: a social critique of the judgement of taste*. London, UK: Routledge.
27. Pareto V. 1971 *Manual of political economy*. London, UK: Oxford University Press.
28. Gastwirth JL. 1972 The estimation of the Lorenz curve and Gini index. *Rev. Econ. Stat.* **54**, 306–316. (doi:10.2307/1937992)
29. The World Bank. *GINI index* (*World Bank estimate*) http://data.worldbank.org/indicator/SI.POV.GINI (accessed 1/2/2016).
30. Souma W. 2000 Physics of personal income. In *Empirical science of financial fluctuations: the advent of econophysics* (ed. H Takayasu), pp. 343–352. Tokyo, Japan: Springer.
31. Aoyama H. 2000 Pareto's law for income of individuals and debt of bankrupt companies. *Fractals* **8**, 293–300.
32. Bottero W. 2005 *Stratification: social division and inequality*. London, UK: Routledge.
33. Mills CW. 1956 *The power elite*. London, UK: Oxford University Press.
34. Zhou S, Mondragón RJ. 2004 The rich-club phenomenon in the internet topology. *IEEE Commun. Lett.* **8**, 180–182. (doi:10.1109/LCOMM.2004.823426)
35. Colizza V, Flammini A, Serrano MA, Vespignani A. 2006 Detecting rich-club ordering in complex networks. *Nat. Phys.* **2**, 110–115. (doi:10.1038/nphys209)
36. Instituto Nacional de Estadística y Geografía. (http://www.inegi.org.mx) 2015 census (accessed 22 February 2016).
37. Wheeler JO. 1967 Occupational status and work-trips: a minimum distance approach. *Soc. Forces* **45**, 508–515. (doi:10.1093/sf/45.4.508)
38. Wheeler JO. 1967 Some effects of occupational status and work trips. *J. Reg. Sci.* **9**, 69–77. (doi:10.1111/j.1467-9787.1969.tb01442.x)
39. Poston DL. 1972 Socioeconomic status and work-residence separation in metropolitan America. *Pac. Sociol. Rev.* **15**, 367–380. (doi:10.2307/1388353)
40. LeRoy S, Sonstelie J. 1983 Paradise lost and regained: transportation innovation, income, and residential location. *J. Urban Econ.* **13**, 67–89. (doi:10.1016/0094-1190(83)90046-3)
41. Rosenthal SS, Ross SL. 2015 Change and persistence in the economic status of neighborhoods and cities. In *Handbook of regional and urban economics*, vol. 5 (eds G Duranton, JV Henderson, WC Strange), pp. 1047–1120. Amsterdam, The Netherlands: Elsevier.
42. Eagle N, Pentland AS, Lazer D. 2009 Inferring friendship network structure by using mobile phone data. *Proc. Natl Acad. Sci. USA* **106**, 15 274–15 278. (doi:10.1073/pnas.0900282106)
43. Šćepanović S, Mishkovski I, Hui P, Nurminen JK, Ylä-Jääski A. 2015 Mobile phone call data as a regional socio-economic proxy indicator. *PLoS ONE* **10**, e0124160. (doi:10.1371/journal.pone.0124160)
44. Blumenstock J, Eagle N. 2010 Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. *ICTD 2010* (*ACM*) **6**, 1–10. (doi:10.1145/2369220.2369225)
45. Mao H, Shuai X, Ahn YY, Bollen J. 2015 Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to Côte d'Ivoire. *EPJ Data Sci.* **4**, 15. (doi:10.1140/epjds/s13688-015-0053-1)
46. Blumenstock J, Cadamuro G, On R. 2015 Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076. (doi:10.1126/science.aac4420)