JOURNAL OF THE ROYAL SOCIETY Interface

rsif.royalsocietypublishing.org

Research

Cite this article: Schläpfer M, Bettencourt LMA, Grauwin S, Raschke M, Claxton R, Smoreda Z, West GB, Ratti C. 2014 The scaling of human interactions with city size. *J. R. Soc. Interface* **11**: 20130789.
http://dx.doi.org/10.1098/rsif.2013.0789

Received: 27 August 2013
Accepted: 6 June 2014

Subject Areas:
mathematical physics, biomathematics

Keywords:
networks, mobile phone data, human interactions, urban scaling, epidemiology

Author for correspondence:
Markus Schläpfer
e-mail: schlmark@mit.edu

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2013.0789 or via http://rsif.royalsocietypublishing.org.

Royal Society Publishing

The scaling of human interactions with city size

Markus Schläpfer[1,2], Luís M. A. Bettencourt[2], Sébastian Grauwin[1], Mathias Raschke[3], Rob Claxton[4], Zbigniew Smoreda[5], Geoffrey B. West[2] and Carlo Ratti[1]

[1]Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Santa Fe Institute, Santa Fe, NM 87501, USA
[3]Raschke Software Engineering, 65195 Wiesbaden, Germany
[4]British Telecommunications PLC, Ipswich IP5 3RE, UK
[5]Orange Labs, 92794 Issy-les-Moulineaux Cedex 9, France

The size of cities is known to play a fundamental role in social and economic life. Yet, its relation to the structure of the underlying network of human interactions has not been investigated empirically in detail. In this paper, we map society-wide communication networks to the urban areas of two European countries. We show that both the total number of contacts and the total communication activity grow superlinearly with city population size, according to well-defined scaling relations and resulting from a multiplicative increase that affects most citizens. Perhaps surprisingly, however, the probability that an individual's contacts are also connected with each other remains largely unaffected. These empirical results predict a systematic and scale-invariant acceleration of interaction-based spreading phenomena as cities get bigger, which is numerically confirmed by applying epidemiological models to the studied networks. Our findings should provide a microscopic basis towards understanding the superlinear increase of different socioeconomic quantities with city size, that applies to almost all urban systems and includes, for instance, the creation of new inventions or the prevalence of certain contagious diseases.

1. Introduction

The statistical relationship between the size of cities and the structure of the network of human interactions at both the individual and population level has so far not been studied empirically in detail. Early-twentieth-century writings suggested that the social life of individuals in larger cities is more fragmented and impersonal than in smaller ones, potentially leading to negative effects such as social disintegration, crime and the development of a number of adverse psychological conditions [1,2]. Although some echoes of this early literature persist today, research since the 1970s has dispelled many of these assumptions by mapping social relations across different places [3,4], yet without providing a comprehensive statistical picture of urban social networks. At the population level, quantitative evidence from many empirical studies points to a systematic acceleration of social and economic life with city size [5,6]. These gains apply to a wide variety of socioeconomic quantities, including economic output, wages, patents, violent crime and the prevalence of certain contagious diseases [7–10]. The average increase in these urban quantities, Y, in relation to the city population size, N, is well described by superlinear scale-invariant laws of the form $Y \propto N^{\beta}$, with a common exponent $\beta \approx 1.15 > 1$ [11,12].

Recent theoretical work suggests that the origin of this superlinear scaling pattern stems directly from the network of human interactions [12–14]—in particular from a similar, scale-invariant increase in social connectivity *per capita* with city size [12]. This is motivated by the fact that human interactions underlie many diverse social phenomena such as the generation of wealth, innovation, crime or the spread of diseases [15–18]. Such conjectures have

not yet been tested empirically, mainly because the measurement of human interaction networks across cities of varying sizes has proved to be difficult to carry out. Traditional methods for capturing social networks—for example through surveys—are time-consuming, necessarily limited in scope, and subject to potential sampling biases [19]. However, the recent availability of many new large-scale datasets, such as those automatically collected from mobile phone networks [20], opens up unprecedented possibilities for the systematic study of urban social dynamics and organization.

In this paper, we explore the relation between city size and the structure of human interaction networks by analysing nationwide communication records in Portugal and the UK. The Portugal dataset contains millions of mobile phone call records collected during 15 months, resulting in an interaction network of 1.6×10^6 nodes and 6.8×10^6 links (reciprocated social ties). In accordance with previous studies on mobile phone networks [21–24], we assume that these nodes represent individuals (subscriptions that indicate business usage are not considered, see Material and methods). Mobile phone communication data are not necessarily a direct representation of the underlying social network. For instance, two individuals may maintain a strong tie through face-to-face interactions or other means of communication, without relying on regular phone calls [23]. Nevertheless, despite such a potential bias, a recent comparison with a questionnaire-based survey has shown that mobile phone communication data are, in general, a reliable proxy for the strength of individual-based social interactions [25]. Moreover, even if two subscribers maintain a close relationship and usually communicate via other means, it seems reasonable to assume that both individuals have called each other at least once during the relatively long observation period of 15 months, thus reducing the chance of missing such relationships in our network [21,26,27]. The UK dataset covers most national landline calls during one month, and the inferred network has 24×10^6 nodes (landline phones) and 119×10^6 links, including reciprocated ties to mobile phones (see Material and methods). We do not consider these nodes as individuals, because we assume that landline phones support the sharing of a single device by several family members or business colleagues [21,28]. Nevertheless, conclusions for the total (i.e. comprising the entire population of a city) social connectivity can be drawn.

With respect to Portugal's mobile phone data, we first demonstrate, that this individual-based interaction network densifies with city size, as the total number of contacts and the total communication activity (call volume and number of calls) grow superlinearly in the number of urban dwellers, in agreement with theoretical predictions and resulting from a continuous shift in the individual-based distributions. Second, we show that the probability that an individual's contacts are also connected with each other (local clustering of links) remains largely constant, which indicates that individuals tend to form tight-knit communities in both small towns and large cities. Third, we show that the empirically observed network densification under constant clustering substantially facilitates interaction-based spreading processes as cities get bigger, supporting the assumption that the increasing social connectivity underlies the superlinear scaling of certain socioeconomic quantities with city size. Additionally, the UK data suggest that the superlinear scaling of the total social connectivity holds for both different means of communication and different national urban systems.

2. Results

2.1. Superlinear scaling of social connectivity

For each city in Portugal, we measured the social connectivity in terms of the total number of mobile phone contacts and the total communication activity (call volume and number of calls). Figure 1*a* shows the total number of contacts (cumulative degree), $K = \sum_{i \in S} k_i$, for each Portuguese city (defined as statistical city, larger urban zone or municipality, see Material and methods) versus its population size, N. Here, k_i is the number of individual i's contacts (nodal degree) and S is the set of nodes assigned to a given city. The variation in K is large, even between cities of similar size, so that a mathematical relationship between K and N is difficult to characterize. However, most of this variation is likely due to the uneven distribution of the telecommunication provider's market share, which for each city can be estimated by the coverage $s = |S|/N$, with $|S|$ being the number of nodes in a given city. While there are large fluctuations in the values of s, we do not find a statistically significant trend with city size that is consistent across all urban units (see the electronic supplementary material). Indeed, rescaling the cumulative degree by s, $K_r = K/s$, substantially reduces its variation (figure 1*b*). Note that this rescaling corresponds to an extrapolation of the observed average nodal degree, $\langle k \rangle = K/|S| = K_r/N$, to the entire city population. Importantly, the relationship between K_r and N is now well characterized by a simple power law, $K_r \propto N^{\beta}$, with exponent $\beta = 1.12 > 1$ (95% confidence interval (CI) [1.11, 1.14]). This superlinear scaling holds over several orders of magnitude and its exponent is in excellent agreement with that of most urban socioeconomic indicators [11] and with theoretical predictions [12]. The small excess of β above unity implies a substantial increase in the level of social interaction with city size: every doubling of a city's population results, on average, in approximately 12% more mobile phone contacts per person, as $\langle k \rangle \propto N^{\beta-1}$ with $\beta - 1 \approx 0.12$. This implies that during the observation period (15 months) an average urban dweller in Lisbon (statistical city, $N = 5.6 \times 10^5$) accumulated about twice as many reciprocated contacts as an average resident of Lixa, a rural town (statistical city, $N = 4.2 \times 10^3$; figure 1*c*). Superlinear scaling with similar values of the exponents also characterizes both the population dependence of the rescaled cumulative call volume, $V_r = \sum_{i \in S} v_i/s$, where v_i is the accumulated time user i spent on the phone, and of the rescaled cumulative number of calls, $W_r = \sum_{i \in S} w_i/s$, where w_i denotes the accumulated number of calls initiated or received by user i (table 1). Together, the similar values of the scaling exponents for both the number of contacts (K_r) and the communication activity (V_r and W_r) also suggest that city size is a less important factor for the weights of links in terms of the call volume and number of calls between each pair of callers. Other city definitions and shorter observation periods [27] lead to similar results with overall $\beta = 1.05-1.15$ (95% CI [1.00, 1.20]). The non-reciprocal (nREC) network (see Material and methods) shows larger scaling exponents $\beta = 1.13-1.24$ (95% CI [1.05, 1.25]), suggesting that the number of social solicitations grows even faster with city size than reciprocated

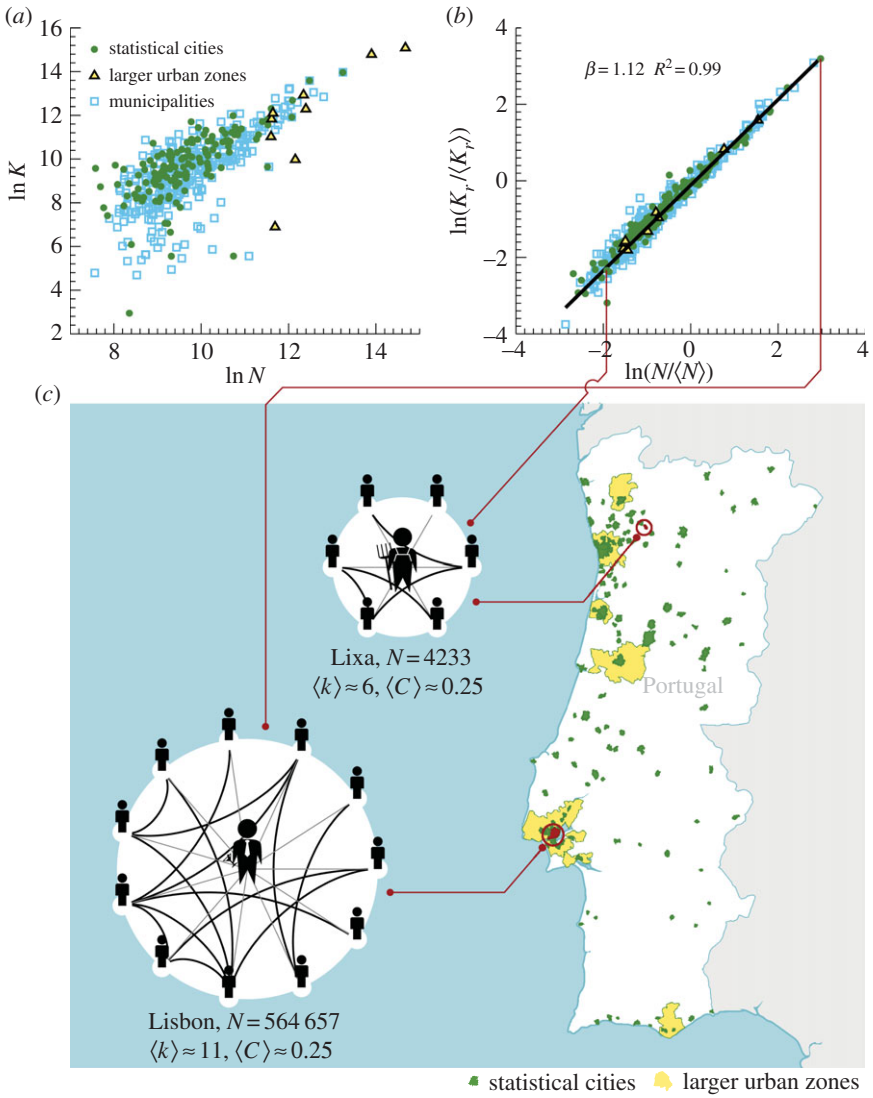

Figure 1. Human interactions scale superlinearly with city size. (*a*) Cumulative degree, K, versus city population size, N, for three different city definitions in Portugal. (*b*) Collapse of the cumulative degree onto a single curve after rescaling by the coverage, $K_r = K/s$. For each city definition, the single values of K_r and N are normalized by their corresponding average values, $\langle K_r \rangle$ and $\langle N \rangle$, for direct comparison across different urban units of analysis. (*c*) An average urban dweller of Lisbon has approximately twice as many reciprocated mobile phone contacts, $\langle k \rangle$, than an average individual in the rural town of Lixa. The fraction of mutually interconnected contacts (black lines) remains unaffected, as indicated by the invariance of the average clustering coefficient, $\langle C \rangle$. The map further depicts the location of the statistical cities and larger urban zones in Portugal, with the exception of those located on the archipelagos of the Azores and Madeira.

contacts. Our predictions for the complete mobile phone coverage are, of course, limited as we observe only a sample of the overall network ($\langle s \rangle \approx 20\%$ for all statistical cities, see Material and methods). Nevertheless, based on the fact that the superlinear scaling also holds when considering only better sampled cities with high values of s (see the electronic supplementary material), and that there is no clear trend in s with city size (so that potential sampling effects presumably apply to urban units of all sizes), we expect that the observed qualitative behaviour also applies to the full network.

For the UK network, despite the relatively short observation period of 31 days, the scaling of reciprocal connectivity shows exponents in the range $\beta = 1.08-1.14$ (95% CI [1.05, 1.17]; table 1). As landline phones may be shared by several people, they do not necessarily reflect an individual-based network, and the meaning of the average degree per device becomes limited. Therefore, and considering that the underlying data cover more than 95% of all residential and business landlines (see Material and methods), we did not rescale the interaction indicators. Nevertheless, the power-law exponents for K, V and W (table 1) support the superlinear scaling of the total social connectivity consistent with Portugal's individual-based network, and suggest that this result applies to both different means of communication and different national urban systems.

2.2. Probability distributions for individual social connectivity

Previous studies of urban scaling have been limited to aggregated, city-wide quantities [11], mainly due to limitations in the availability and analysis of extensive individual-based data covering entire urban systems. Here, we leverage the granularity of our data to explore how scaling relations emerge from the underlying distributions of network properties. We focus on Portugal as, in comparison with landlines, mobile phone communication provides a more direct proxy for person-to-person interactions [25,29,30] and is generally known to correlate well with other means of communication [21] and face-to-face meetings [31]. Moreover, for this part of our analysis, we considered only regularly active callers who

Table 1. Scaling exponents β. The observation period of $\Delta T = 409$ days is the full extent of the Portugal dataset, while $\Delta T = 92$ days corresponds to the first three consecutive months. For the call volume statistics, we discarded one larger urban zone (Ponta Delgada) due to a high estimation error of V_r (s.e.m. $> 20\%$). For the UK data, the interaction indicators, Y, are not rescaled by the coverage due to the consistently high market share of the telecommunication provider. The indicator K_{lm} is based on the cumulative number of links between landlines and mobile phones only (landline–landline connections are excluded). Exponents were estimated by nonlinear least-squares regression (trust-region algorithm), with adj.-$R^2 > 0.98$ for all fits.

city definition	number	network type	ΔT (days)	Y	β	95% CI
Portugal						
statistical city	140	reciprocal	409	degree (K_r)	1.12	[1.11, 1.14]
				call volume (V_r)	1.11	[1.09, 1.12]
				number of calls (W_r)	1.10	[1.09, 1.11]
			92	degree (K_r)	1.10	[1.09, 1.11]
				call volume (V_r)	1.10	[1.08, 1.11]
				number of calls (W_r)	1.08	[1.07, 1.10]
		non-reciprocal	409	degree (K_r)	1.24	[1.22, 1.25]
				call volume (V_r)	1.14	[1.12, 1.15]
				number of calls (W_r)	1.13	[1.12, 1.14]
larger urban zone	9(8)	reciprocal	409	degree (K_r)	1.05	[1.00, 1.11]
				call volume (V_r)	1.11	[1.02, 1.20]
				number of calls (W_r)	1.10	[1.05, 1.15]
		non-reciprocal	409	degree (K_r)	1.13	[1.08, 1.18]
				call volume (V_r)	1.14	[1.05, 1.23]
				number of calls (W_r)	1.13	[1.08, 1.18]
municipality	293	reciprocal	409	degree (K_r)	1.13	[1.11, 1.14]
				call volume (V_r)	1.15	[1.13, 1.17]
				number of calls (W_r)	1.13	[1.11, 1.14]
UK						
urban audit city	24	reciprocal	31	degree (K)	1.08	[1.05, 1.12]
				degree, land-mobile (K_{lm})	1.14	[1.11, 1.17]
				call volume (V)	1.10	[1.07, 1.14]
				number of calls (W)	1.08	[1.05, 1.11]

initiated and received at least one call during each successive period of three months, so as to avoid a potential bias towards longer periods of inactivity (see the electronic supplementary material). The resulting statistical distributions of the nodal degree, call volume and number of calls are remarkably regular across diverse urban settings, with a clear shift towards higher values with increasing city size (figure 2).

To estimate the type of parametric probability distribution that best describes these data, we selected as trial models (i) the lognormal distribution, (ii) the generalized Pareto distribution, (iii) the double Pareto-lognormal distribution and (iv) the skewed lognormal distribution (see the electronic supplementary material). We first calculated for each interaction indicator, each model i and individual city c the maximum value of the log-likelihood function $\ln L_{i,c}$ [32]. We then deployed it to quantify the Bayesian information criterion (BIC) as $\mathrm{BIC}_{i,c} = -2 \ln L_{i,c} + \eta_i |S_c|$, where η_i is the number of parameters used in model i and $|S_c|$ is the sample size (number of callers in city c). The model with the lowest BIC is selected as the best model (see the electronic supplementary material, tables S7–S9). We find that the statistics of the nodal degree is well described by a skewed lognormal distribution (i.e. $k^* = \ln k$ follows a skew-normal distribution), whereas both the call volume and the number of calls are well approximated by a conventional lognormal distribution (i.e. $v^* = \ln v$ and $w^* = \ln w$ follow a Gaussian distribution). The mean values of all logarithmic variables are consistently increasing with city size (figure 2, insets). While there are some trends in the standard deviations (e.g. the standard deviation of k^* is slightly increasing for the municipalities and the standard deviation of v^* is decreasing for the statistical cities), overall, we do not observe a clear behaviour consistent across all city definitions. This indicates that superlinear scaling is not simply due to the dominant effect of a few individuals (as in a power-law distribution), but results from an increase in the individual connectivity that characterizes most callers in the city.

More generally, lognormal distributions typically appear as the limit of many random multiplicative processes [33], suggesting that an adequate model for the generation of new acquaintances would need to consider a stochastic cascade of new social encounters in space and time that is facilitated in larger cities. As for the analysis of the city-wide quantities (section 2.1), the average coverage of $\langle s \rangle \approx 20\%$ may limit our prediction for the complete communication network due

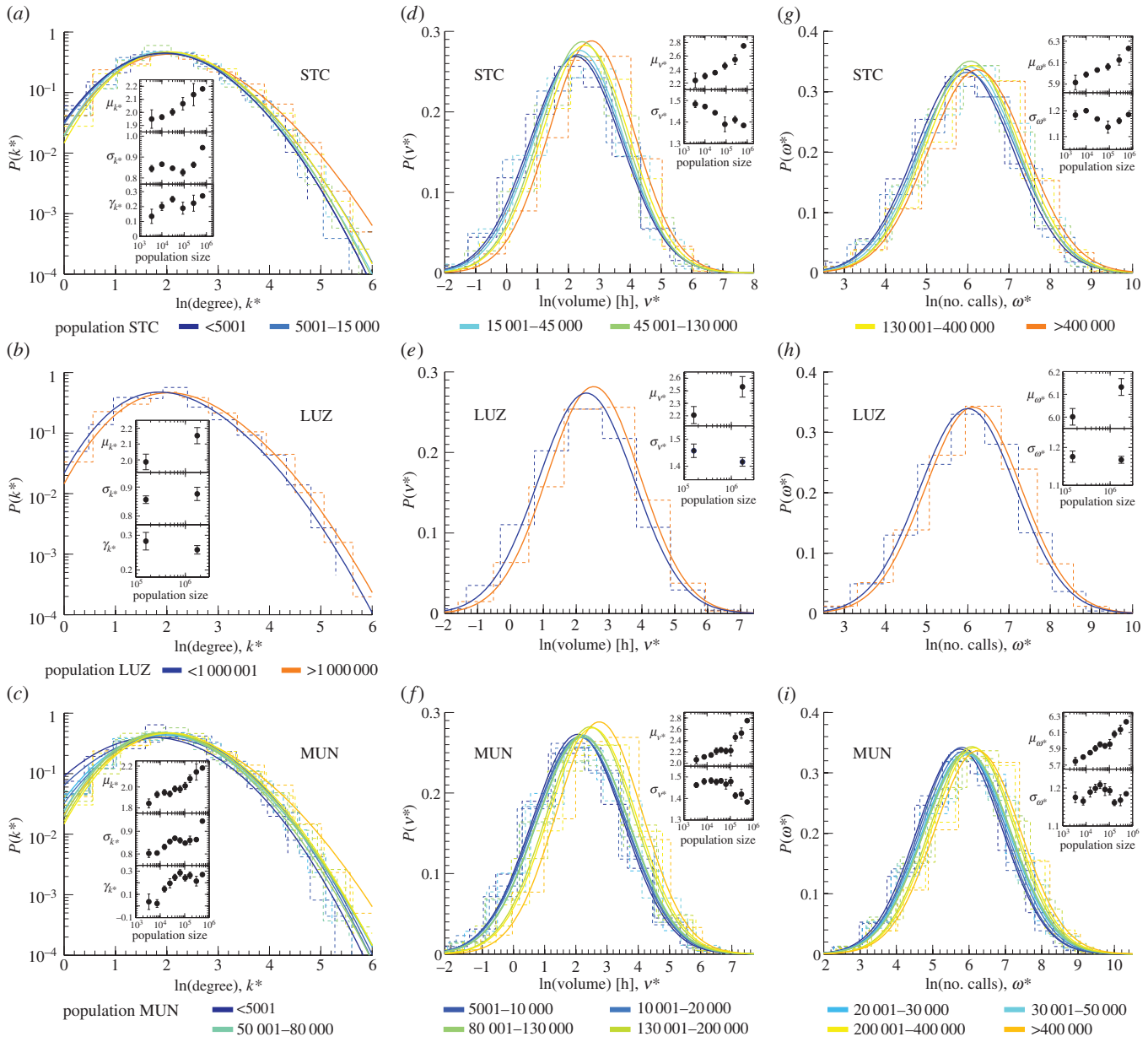

Figure 2. The impact of city size on human interactions at the individual level. (*a–c*) Degree distributions, $P(k^*)$, for statistical cities (STC), larger urban zones (LUZ) and municipalities (MUN); the individual urban units are log-binned according to their population size. The dashed lines indicate the underlying histograms and the continuous lines are best fits of the skew-normal distribution with mean μ_{k^*}, standard deviation σ_{k^*} and skewness γ_{k^*} (insets). (*d–f*) Distributions of the call volume, $P(v^*)$ and (*g–i*) number of calls, $P(w^*)$; the continuous lines are best fits of the normal distribution with mean values μ_{v^*} and μ_{w^*} and standard deviations σ_{v^*} and σ_{w^*}, respectively (insets). Error bars denote the standard error of the mean (s.e.m.). The distribution parameters are estimated by the maximum-likelihood method, see the electronic supplementary material.

to potential sampling effects [34,35]. However, as the basic shape of the distributions is preserved even for those cities with a very high coverage (see the electronic supplementary material, figure S6), we hypothesize that the observed qualitative behaviour also holds for $\langle s \rangle \approx 100\%$.

2.3. Invariance of the average clustering coefficient

Finally, we examined the local clustering coefficient, C_i, which measures the fraction of connections between one's social contacts to all possible connections between them [36]; that is $C_i \equiv 2z_i/[k_i(k_i-1)]$, where z_i is the total number of links between the k_i neighbours of node i. A high value of C_i (close to unity) indicates that most of one's contacts also know each other, whereas if $C_i = 0$, they are mutual strangers. As larger cities provide a larger pool from which contacts can be selected, the probability that two contacts are also mutually connected would decrease rapidly if they were established at random (see the electronic supplementary material). In contrast to this expectation, we find that the clustering coefficient averaged over all nodes in a given city, $\langle C \rangle = \sum_{i \in S} C_i/|S|$, remains approximately constant with $\langle C \rangle \approx 0.25$ in the individual-based network in Portugal (figures 1*c* and 3). Moreover, the clustering remains largely unaffected by city size, even when taking into account the link weights (call volume and number of calls, see the electronic supplementary material). The fact that we observe only a sample of the overall mobile phone network in Portugal may have an influence on the absolute value of $\langle C \rangle$ [35], especially if tight social groups may prefer using the same telecommunication provider. Nevertheless, we expect that this potential bias has no effect on the invariance

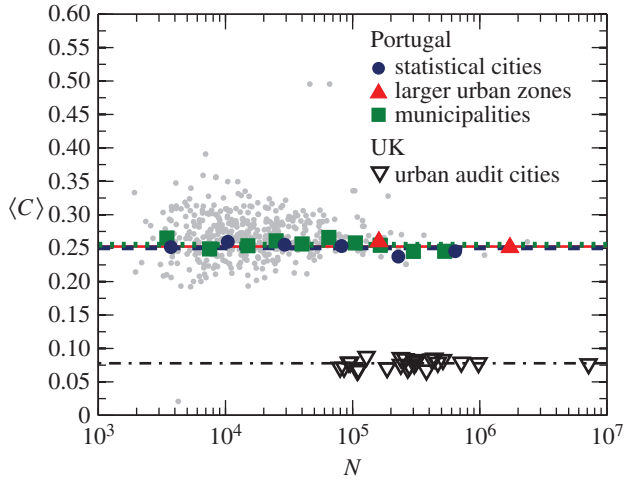

Figure 3. The average clustering coefficient remains unaffected by city size. The lines indicate the average values with 0.251 ± 0.021 for STC (weighted average and standard deviation, dashed line), 0.252 ± 0.013 for LUZ (continuous line) and 0.255 ± 0.021 for MUN (dotted line) in Portugal, and 0.078 ± 0.004 for the UK (dashed-dotted line). For Portugal, the individual urban units are log-binned according to their population size as in figure 2, to compensate for the varying coverage of the telecommunication provider. The error bars (s.e.m.) are smaller than the symbols. Grey points are the underlying scatter plot for all urban units. A regression analysis on the data is provided in the electronic supplementary material, figure S7. The values of $\langle C \rangle$ in the UK are lower than those in Portugal, as expected for a landline network that captures the aggregated activity of different household members or business colleagues. If we assume that an average landline in the UK is used by three people who communicate with a separate set of unconnected friends, we would indeed expect that the clustering coefficient would be approximately one-third of that of each individual.

of $\langle C \rangle$, as we do not find a clear trend in the coverage s with city size (see the electronic supplementary material). Thus, assuming that the analysed mobile phone data are a reliable proxy for the strength of social relations [25], the constancy of the average clustering coefficient with city size indicates, perhaps surprisingly, that urban social networks retain much of their local structures as cities grow, while reaching further into larger populations. In this context, it is worth noting that the mobile phone network in Portugal exhibits assortative degree–degree correlations, denoting the tendency of a node to connect to other nodes with similar degree [37] (see the electronic supplementary material). The presence of assortative degree–degree correlations in networks is known to allow high levels of clustering [38].

2.4. Acceleration of spreading processes

The empirical quantities analysed so far are topological key factors for the efficiency of network-based spreading processes, such as the diffusion of information and ideas or the transmission of diseases [39]. The degree and communication activity (call volume and number of calls) indicate how fast the state of a node may spread to nearby nodes [15,40,41], whereas the clustering largely determines its probability of propagating beyond the immediate neighbours [42,43]. Hence, considering the invariance of the link clustering, the connectivity increase (table 1) suggests that individuals living in larger cities tend to have similar, scale-invariant gains in their spreading potential compared with those living in smaller towns. Given the continuous shift of the underlying distributions (figure 2), this increasing influence seems to involve most urban dwellers. However, several non-trivial network effects such as community structures [24] or assortative mixing by degree [44] may additionally play a crucial role in the resulting spreading dynamics.

Thus, to directly test whether the increasing connectivity implies an acceleration of spreading processes, we applied a simple epidemiological model to Portugal's individual-based mobile phone network. The model has been introduced in reference [21] for the analysis of information propagation through mobile phone communication, and is similar to the widely used susceptible–infected model in which the nodes are either in a susceptible or infected state [15]. The spreading is captured by the dynamic state variable $\xi_i(t) \in \{0, 1\}$ assigned to each node i, with $\xi_i(t) = 1$ if the node is infected (or informed) and $\xi_i(t) = 0$ otherwise. For a given city c, we set at time $t = 0$ the state of a randomly selected node $i \in S_c$ to $s_i(0) = 1$, whereas all other nodes are in the susceptible (or not-informed) state. At each subsequent time step, an infected node i can pass the information on to each susceptible nearest neighbour j with probability $P_{ij} = x v_{ij}$, where v_{ij} is the weight of the link between node i and node j in terms of the accumulated call volume, and the parameter x determines the overall spreading speed. Hence, the chance that two individuals will communicate the information increases with the accumulated time they spend on the phone. In accordance with reference [21], we choose $x = 1/v_{0.9} = 1/6242\ \mathrm{s}^{-1}$, with $v_{0.9}$ being the value below which 90% of all link weights in the network fall. This threshold allows reduction of the problem of long simulation running times owing to the broad distribution of the link weights, whereas $P_{ij} \propto v_{ij}$ holds for 90% of all links in the network. The propagation is always realized for the strongest 10% of the links ($P_{ij} = 1$, see [21]). For each simulation run κ, we measured the time $t_{c,\kappa}(n_I)$ until $n_I = \sum_{i \in S_c} \xi_i(t)$ nodes in the given city were infected and estimated the spreading speed as $R_{c,\kappa} = n_I / t_{c,\kappa}(n_I)$. The average spreading speed for city c is then given by averaging over all simulation runs, $R_c = \langle R_{c,\kappa} \rangle$. The spreading paths are not restricted to city boundaries, but may involve the entire nationwide network. We set the total number of infected nodes to $n_I = 100$ and discarded four statistical cities and 17 municipalities for which $|S| < n_I$. Examples for the infection dynamics and the distribution of the spreading speed resulting from single runs are provided in the electronic supplementary material, figure S10. Figure 4 depicts the resulting values of R for all cities. Indeed, we find a systematic increase of the spreading speed with city size, that can again be approximated by a power-law scaling relation, $R \propto N^{\delta}$, with $\delta = 0.11 - 0.15$ (95% CI [0.02, 0.26]). Similar increases are also found for simulations performed on the unweighted network (see the electronic supplementary material, figure S11). These numerical results thus confirm the expected acceleration of spreading processes with city size, and are also in line with a recent simulation study on synthetic networks [14]. Moreover, such an increase in the spreading speed is considered to be a key ingredient for the explanation of the superlinear scaling of certain socioeconomic quantities with city size [12,14] as, for instance, rapid information diffusion and the efficient exchange of ideas over person-to-person networks can be linked to innovation and productivity [12,45].

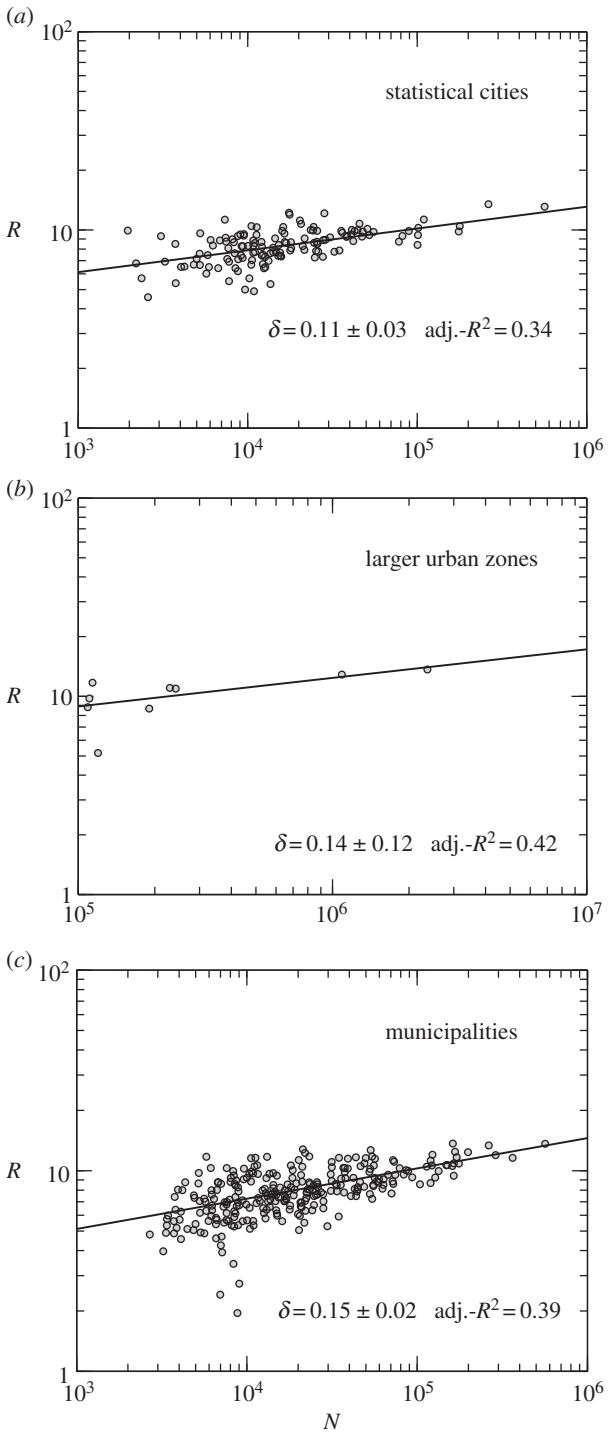

Figure 4. Larger cities facilitate interaction-based spreading processes. The panels show the average spreading speed versus city size, broken down into the different city definitions. For each urban unit, the values of R result from averaging over 100 simulation trials performed on the reciprocal network in Portugal ($\Delta T = 409$ days), weighted by the accumulated call volume between each pair of nodes. The solid lines are the best fit of a power-law scaling relation $R \propto N^{\delta}$, for which the values of the exponent, the corresponding 95% CIs and the coefficients of determination are indicated.

3. Discussion

By mapping society-wide communication networks to the urban areas of two European countries, we were able to empirically test the hypothesized scale-invariant increase of human interactions with city size. The observed increase is substantial and takes place within well-defined behavioural constraints in that (i) the total number of contacts (degree) and the total communication activity (call volume and number of calls) obey superlinear power-law scaling in agreement with theory [12] and resulting from a multiplicative increase that affects most citizens, whereas (ii) the average local clustering coefficient does not change with city size. Assuming that the analysed data are a reasonable proxy for the strength of the underlying social relations [25], and that our results apply to the complete interaction networks, the constant clustering is particularly noteworthy as it suggests that even in large cities we live in groups that are as tightly knit as those in small towns or 'villages' [46]. However, in a real village, we may need to accept a community imposed on us by sheer proximity, whereas in a city, we can follow the homophilic tendency [47] of choosing our own village—people with shared interests, profession, ethnicity, sexual orientation, etc. Together, these characteristics of the analysed communication networks indicate that larger cities may facilitate the diffusion of information and ideas or other interaction-based spreading processes. This further supports the prevailing hypothesis that the structure of social networks underlies the generic properties of cities, manifested in the superlinear scaling of almost all socioeconomic quantities with population size.

The wider generality of our results remains, of course, to be tested on other individual-based communication data, ideally with complete coverage of the population ($\langle s \rangle \approx 100\%$). Nevertheless, the revealed patterns offer a baseline to additionally explore the differences of particular cities with similar size, to compare the observed network properties with face-to-face interactions [31] and to extend our study to other cultures and economies. Furthermore, it would be instructive to analyse in greater detail how cities affect more specific circles of social contacts such as family, friends or business colleagues [22,25]. Finally, it remains a challenge for future studies to establish the causal relationship between social connectivity at the individual and organizational levels and the socioeconomic characteristics of cities, such as economic output, the rate of new innovations, crime or the prevalence of contagious diseases. To that end, in combination with other socioeconomic or health-related data, our findings might serve as a microscopic and statistical basis for network-based interaction models in sociology [20,48], economics [7,49] and epidemiology [18].

4. Material and methods

4.1. Datasets

The Portugal dataset consists of 440 million call detail records (CDRs) from 2006 and 2007, covering voice calls of ≈2 million mobile phone users and thus ≈20% of the country's population (in 2006, the total mobile phone penetration rate was ≈100%, survey available at http://www.anacom.pt). The data have been collected by a single telecom service provider for billing and operational purposes. The overall observation period is 15 months during which the data from 46 consecutive days are lacking, resulting in an effective analysis period of $\Delta T = 409$ days. To safeguard privacy, individual phone numbers were anonymized by the operator and replaced with a unique security ID. Each CDR consists of the IDs of the two connected individuals, the call duration, the date and time of the call initiation, as well as the unique IDs of the two cell towers routing the call at its initiation. In total, there are 6511 cell towers for which the geographical location was provided, each serving on average an area of 14 km^2, which reduces to 0.13 km^2 in urban

areas. The UK dataset contains 7.6 billion calls from a one-month period in 2005, involving 44 million landline and 56 million mobile phone numbers (greater than 95% of all residential and business landlines countrywide). For customer anonymity, each number was replaced with a random, surrogate ID by the operator before providing the data. We had only partial access to the connections made between any two mobile phones. The operator partitioned the country into 5500 exchange areas (covering 49 km^2 on average), each of which comprises a set of landline numbers. The dataset contains the geographical location of 4000 exchange areas.

4.2. City definitions

Because there is no unambiguous definition of a city we explored different units of analysis. For Portugal, we used the following city definitions: (i) statistical cities (STC), (ii) municipalities (MUN) and (iii) larger urban zones (LUZ). STC and MUN are defined by the Portuguese National Statistics Office (http://www.ine.pt), which provided us with the 2001 population data, and with the city perimeters (shapefiles containing spatial polygons). The LUZ are defined by the European Union Statistical Agency (Eurostat) and correspond to extended urban regions (the population statistics and shapefiles are publicly available at http://www.urbanaudit.org). For the LUZ, we compiled the population data for 2001 to assure comparability with the STC and MUN. In total, there are 156 STC, 308 MUN and nine LUZ. The MUN are an administrative subdivision and partition the entire national territory. Although their interpretation as urban units is flawed in some cases, the MUN were included in the study as they cover the total resident population of Portugal. There are six MUN which correspond to an STC. For the UK, we focused on urban audit cities (UACs) as defined by Eurostat, being equivalent to local administrative units, level 1 (LAU-1). Thus, using population statistics for 2001 allows for a direct comparison with the MUN in Portugal (corresponding to LAU-1). In total, the UK contains 30 UAC.

4.3. Spatial interaction networks

For Portugal, we inferred two distinct types of interaction networks from the CDRs: in the reciprocal (REC) network, each node represents a mobile phone user, and two nodes are connected by an undirected link if each of the two corresponding users initiated at least one call to the other. In accordance with previous studies on mobile phone data [21,22], this restriction to reciprocated links avoids subscriptions that indicate business usage (large number of calls which are never returned) and should largely eliminate call centres or accidental calls to wrong subscribers. In the nREC network, two nodes are connected if there has been at least one call between them. The nREC network thus contains one-way calls that were never reciprocated, presumably representing more superficial interactions between individuals who might not know each other personally. Nevertheless, we eliminated all nodes which never received or never initiated any call, so as to avoid a potential bias induced by call centres and other business hubs. We performed our study on the largest connected component (LCC, corresponding to the giant weakly connected component for the nREC network) extracted from both network types (see the electronic supplementary material, table S1). In order to assign a given user to one of the different cities, we first determined the cell tower which routed most of his/her calls, presumably representing his or her home place. Subsequently, the corresponding geographical coordinate pairs were mapped to the polygons (shapefiles) of the different cities. Following this assignment procedure, we were left with 140 STC (we discarded five STC for which no shapefile was available and 11 STC without any assigned cell tower), nine LUZ and 293 MUN (we discarded 15 MUN without any assigned cell tower), see the electronic supplementary material, figure S1 and table S2, for the population statistics. The number of assigned nodes is strongly correlated with city population size ($r = 0.95$, 0.97, 0.92 for STC, LUZ and MUN, respectively, with p-value < 0.0001 for the different urban units), confirming the validity of the applied assignment procedure. To further test the robustness of our results, we additionally determined the home cell tower by considering only those calls that were initiated between 22.00 and 07.00, yielding qualitatively similar findings to those reported in the main text. For the UK, owing to limited access to calls among mobile phones and to insufficient information about their spatial location, we included only those mobile phone numbers that had at least one connection to a landline phone. Subsequently, in order to reduce a potential bias induced by business hubs, we followed the data-filtering procedure used in [49]. Hence, we considered only the REC network and we excluded all nodes with a degree larger than 50, as well as all links with a call volume exceeding the maximum value observed for those links involving mobile phone users. Summary statistics are given in the electronic supplementary material, table S3. We then assigned an exchange area together with its set of landline numbers to a UAC, if the centre point of the former is located within the polygon of the latter. This results in 24 UAC containing at least one exchange area (see the electronic supplementary material, figure S2 and table S4).

Acknowledgements. We thank José Lobo, Stanislav Sobolevsky, Michael Szell, Riccardo Campari, Benedikt Gross and Janet Owers for comments and discussions. M.S., S.G. and C.R. gratefully acknowledge British Telecommunications PLC, Orange Labs, the National Science Foundation, the AT&T Foundation, the MIT Smart Programme, Ericsson, BBVA, GE, Audi Volkswagen, Ferrovial and the members of the MIT Senseable City Laboratory Consortium where this work was carried out as part of Ericsson's "Signature of Humanity" programme.

Funding statement. L.M.A.B. and G.B.W. acknowledge partial support by the Rockefeller Foundation, the James S. McDonnell Foundation (grant no. 220020195), the National Science Foundation (grant no. 103522), the John Templeton Foundation (grant no. 15705) and the Army Research Office Minerva Programme (grant no. W911NF1210097). Mobile phone and landline data were provided by anonymous service providers in Portugal and the UK and are not available for distribution.

References

1. Simmel G. 1950 *The sociology of Georg Simmel* (trans. and ed. Wolff KH). New York, NY: Free Press.
2. Wirth L. 1938 Urbanism as a way of life. *Am. J. Sociol.* **44**, 1–24. (doi:10.1086/217913)
3. Fischer CS. 1982 *To dwell among friends: personal networks in town and country*. Chicago, IL: University of Chicago Press.
4. Wellman B. 1999 *Networks in the global village: life in contemporary communities*. Boulder, CO: Westview Press.
5. Milgram S. 1970 The experience of living in cities. *Science* **167**, 1461–1468. (doi:10.1126/science.167.3924.1461)
6. Bornstein MH, Bornstein HG. 1976 The pace of life. *Nature* **259**, 557–559. (doi:10.1038/259557a0)
7. Fujita M, Krugman P, Venables AJ. 2001 *The spatial economy: cities, regions, and international trade*. Cambridge, MA: MIT Press.
8. Sveikauskas L. 1975 The productivity of cities. *Q. J. Econ.* **89**, 393–413. (doi:10.2307/1885259)
9. Cullen JB, Levitt SD. 1999 Crime, urban flight, and the consequences for cities. *Rev. Econ. Stat.* **81**, 159–169. (doi:10.1162/003465399558030)

10. Centers for Disease Control and Prevention. 2012 *HIV surveillance in urban and nonurban areas*. See http://www.cdc.gov.
11. Bettencourt LMA, Lobo J, Helbing D, Kühnert C, West GB. 2007 Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl Acad. Sci. USA* **104**, 7301–7306. (doi:10.1073/pnas.0610172104)
12. Bettencourt LMA. 2013 The origin of scaling in cities. *Science* **340**, 1438–1441. (doi:10.1126/science.1235823)
13. Arbesman S, Kleinberg JM, Strogatz SH. 2009 Superlinear scaling for innovation in cities. *Phys. Rev. E* **79**, 016115. (doi:10.1103/PhysRevE.79.016115)
14. Pan W, Ghoshal G, Krumme C, Cebrian M, Pentland A. 2013 Urban characteristics attributable to density-driven tie formation. *Nat. Commun.* **4**, 1961. (doi:10.1038/ncomms2961)
15. Anderson RM, May RM. 1991 *Infectious diseases of humans: dynamics and control*. Oxford, UK: Oxford University Press.
16. Rogers EM. 1995 *Diffusion of innovation*. New York, NY: Free Press.
17. Topa G. 2001 Social interactions, local spillovers and unemployment. *Rev. Econ. Stud.* **68**, 261–295. (doi:10.1111/1467-937X.00169)
18. Eubank S, Guclu H, Kumar VSA, Marathe MV, Srinivasan A, Toroczkai Z, Wang N. 2004 Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180–184. (doi:10.1038/nature02541)
19. Berk RA. 1983 An introduction to sample selection bias in sociological data. *Am. Sociol. Rev.* **48**, 386–398. (doi:10.2307/2095230)
20. Lazer D *et al*. 2009 Computational social science. *Science* **323**, 721–723. (doi:10.1126/science.1167742)
21. Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabási A-L. 2007 Structure and tie strength in mobile communication networks. *Proc. Natl Acad. Sci. USA* **104**, 7332–7336. (doi:10.1073/pnas.0610245104)
22. Miritello G, Lara R, Cebrian M, Moro E. 2013 Limited communication capacity unveils strategies for human interaction. *Sci. Rep.* **3**, 1950. (doi:10.1038/srep01950)
23. Raeder T, Lizardo Chawla NV, Hachen D. 2011 Predictors of short-term deactivation of cell phone contacts in a large scale communication network. *Soc. Net.* **33**, 245–257. (doi:10.1016/j.socnet.2011.07.002)
24. Karsai M, Kivelä M, Pan RK, Kaski K, Kertész J, Barabási A-L, Saramäki J. 2011 Small but slow world: how network topology and burstiness slow down spreading. *Phys. Rev. E* **83**, 025102. (doi:10.1103/PhysRevE.83.025102)
25. Saramäki J, Leicht EA, López E, Roberts SGB, Reed-Tsochas F, Dunbar RIM. 2014 Persistence of social signatures in human communication. *Proc. Natl Acad. Sci. USA* **111**, 942–947. (doi:10.1073/pnas.1308540110)
26. Licoppe C, Smoreda Z. 2005 Are social networks technically embedded? How networks are changing today with changes in communication technology. *Soc. Net.* **27**, 317–335. (doi:10.1016/j.socnet.2004.11.001)
27. Krings G, Karsai M, Bernhardsson S, Blondel VD, Saramäki J. 2012 Effects of time window size and placement on the structure of aggregated networks. *EPJ Data Sci.* **1**, 1–16. (doi:10.1140/epjds4)
28. Geser H. 2006 Is the cell phone undermining the social order? Understanding mobile technology from a sociological perspective. *Know. Technol. Pol.* **19**, 8–18. (doi:10.1007/s12130-006-1010-x)
29. Eagle N, Pentland A, Lazer D. 2009 Inferring friendship structure by using mobile phone data. *Proc. Natl Acad. Sci. USA* **106**, 15 274–15 278. (doi:10.1073/pnas.0900282106)
30. Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO. 2013 The impact of biases in mobile phone ownership on estimates of human mobility. *J. R. Soc. Interface* **10**, 20120986. (doi:10.1098/rsif.2012.0986)
31. Calabrese F, Smoreda Z, Blondel VD, Ratti C. 2011 Interplay between telecommunications and face-to-face interactions: a study using mobile phone data. *PLoS ONE* **6**, e20814. (doi:10.1371/journal.pone.0020814)
32. Davidson AC. 2003 *Statistical models*. Cambridge, UK: Cambridge University Press.
33. Mitzenmacher M. 2004 A brief history of generative models for power law and lognormal distributions. *Internet Math.* **1**, 226–251. (doi:10.1080/15427951.2004.10129088)
34. Stumpf MPH, Wiuf C, May RM. 2005 Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl Acad. Sci. USA* **102**, 4221–4224. (doi:10.1073/pnas.0501179102)
35. Lee SH, Kim PJ, Jeong H. 2006 Statistical properties of sampled networks. *Phys. Rev. E* **73**, 016102. (doi:10.1103/PhysRevE.73.016102)
36. Watts DJ, Strogatz SH. 1998 Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442. (doi:10.1038/30918)
37. Raschke M, Schläpfer M, Nibali R. 2010 Measuring degree–degree association in networks. *Phys. Rev. E* **82**, 037102. (doi:10.1103/PhysRevE.82.037102)
38. Serrano MA, Boguña M. 2005 Tuning clustering in random networks with arbitrary degree distributions. *Phys. Rev. E* **72**, 036133. (doi:10.1103/PhysRevE.72.036133)
39. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D. 2006 Complex networks: structure and dynamics. *Phys. Rep.* **424**, 175–308. (doi:10.1016/j.physrep.2005.10.009)
40. Pastor-Satorras R, Vespignani A. 2001 Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200–3203. (doi:10.1103/PhysRevLett.86.3200)
41. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA. 2010 Identification of influential spreaders in complex networks. *Nat. Phys.* **6**, 888–893. (doi:10.1038/nphys1746)
42. Newman MEJ. 2009 Random graphs with clustering. *Phys. Rev. Lett.* **103**, 058701. (doi:10.1103/PhysRevLett.103.058701)
43. Granovetter M. 1973 The strength of weak ties. *Am. J. Sociol.* **78**, 1360–1380. (doi:10.1086/225469)
44. Kiss IZ, Green DM, Kao RR. 2008 The effect of network mixing patterns on epidemic dynamics and the efficacy of disease contact tracing. *J. R. Soc. Interface* **5**, 791–799. (doi:10.1098/rsif.2007.1272)
45. Granovetter M. 2005 The impact of social structure on economic outcomes. *J. Econ. Persp.* **19**, 33–50. (doi:10.1257/0895330053147958)
46. Jacobs J. 1961 *The death and life of great American cities*. New York, NY: Random House.
47. 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)
48. Wasserman S, Faust K. 1994 *Social network analysis: methods and applications*. Cambridge, UK: Cambridge University Press.
49. Eagle N, Macy M, Claxton R. 2010 Network diversity and economic development. *Science* **328**, 1029–1031. (doi:10.1126/science.1186605)