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, USA*

[2]*Santa Fe Institute, Santa Fe, NM, USA*

[3]*Raschke Software Engineering, Wiesbaden, Germany*

[4]*British Telecommunications plc, Ipswich, UK*

[5]*Orange Labs, Issy-les-Moulineaux, France*

Abstract

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.

I. 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 20th 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 proven 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 data sets, such as those automatically collected from mobile phone networks [20], opens up unprecedented possibilities for the systematic study of the urban social dynamics and organisation.

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 data set 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 (recipro-

cated 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 data set covers most national landline calls during 1 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 demonstrate first, 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.

II. RESULTS

A. 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 characterise. 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 characterised 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, $N = 5 \times 10^5$) accumulated about twice as many reciprocated contacts as an average resident of Lixa, a rural town (Statistical City, $N = 4 \times 10^3$, see figure 1*c*). Superlinear scaling with similar values of the exponents also characterises 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, see 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 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 contacts. Our predictions for the complete mobile phone coverage are, of course, limited as we only observe 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]), see 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 covers 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.

B. 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 to 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 initiated and received at least one call during each successive period of 3 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 generalised 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), while 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 characterises 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 I.A), the average coverage of $\langle s \rangle \approx 20\%$ may limit our prediction for the complete communication network due 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 hypothesise that the observed qualitative behaviour also holds for $\langle s \rangle \approx 100\%$.

C. 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, while 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 (figure 1*c* and figure 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 only observe 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 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].

D. 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], while 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 to 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 ref. [21] for the analysis of information propagation through mobile phone communication, and is similar to the widely used susceptible-infected (SI) 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 $s_i(0) = 1$, while 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 $P_{ij} = x\nu_{ij}$, where ν_{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 ref. [21], we choose $x = 1/\nu_{0.9} = 1/6242s^{-1}$, with $\nu_{0.9}$ being the value below which 90% of all link

weights in the network fall. This threshold allows to reduce the problem of long simulation running times due to the broad distribution of the link weights, while $P_{ij} \propto \nu_{ij}$ holds for 90% of all links in the network. The propagation is always realised 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 4 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 has been 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].

III. DISCUSSION

By mapping society-wide communication networks to the urban areas of two European countries, we were able to empirically test the hypothesised 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, number of calls) obey superlinear power-law scaling in agreement with theory [12] and resulting from a multiplicative increase that affects most citizens, while *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 organisational 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].

IV. MATERIAL AND METHODS

A. Data sets

The Portugal data set consists of 440 million Call Detail Records (CDR) from 2006 and 2007, covering voice calls of ≈ 2 million mobile phone users and thus $\approx 20\%$ of the country's population (in 2006 the total mobile phone penetration rate was $\approx 100\%$, survey available at http://www.anacom.pt). The data has 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 is lacking, resulting in an effective analysis period of $\Delta T = 409$ days. To safeguard privacy, individual phone numbers were anonymised 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 geographic 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 data set contains 7.6 billion calls from a one-month period in 2005, involving 44 million landline and 56 million mobile phone numbers ($> 95\%$ 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 data set contains the geographic location of 4000 exchange areas.

B. 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 9 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 6 MUN which correspond to a STC. For the UK, we focussed on Urban Audit Cities (UAC) 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.

C. 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 centers or accidental calls to wrong subscribers. In the non-reciprocal (nREC) network two nodes are connected if there has been at least one call between them. The nREC network thus contains one-way calls which were never reciprocated, presumably representing more superficial interactions between individuals which 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 cluster (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 geographic coordinate pairs were mapped to the polygons (shapefiles) of the different cities. Following this assignment procedure, we were left with 140 STC (we discarded 5 STC for which no shapefile was available and 11 STC without any assigned cell tower), 9 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 10pm and 7am, yielding qualitatively similar findings to those reported in the main text. For the

UK, due 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 an 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).

---

[1] Simmel G. 1950 *The sociology of Georg Simmel* (trans. and ed. Wolff KH). New York: Free Press.

[2] Wirth L. 1938 Urbanism as a way of life. *Am. J. Sociol.* **44,** 1-24.

[3] Fischer CS. 1982 *To dwell among friends: personal networks in town and country.* Chicago: University of Chicago Press.

[4] Wellman B. 1999 *Networks in the global village: life in contemporary communities.* Boulder, USA: Westview Press.

[5] Milgram S. 1970 The experience of living in cities. *Science* **167,** 1461-1468.

[6] Bornstein MH, Bornstein HG. 1976 The pace of life. *Nature* **259,** 557-559.

[7] Fujita M, Krugman P, Venables AJ. 2001 *The spatial economy: cities, regions, and international trade.* Cambridge, USA: MIT Press.

[8] Sveikauskas L. 1975 The productivity of cities. *Q. J. Econ.* **89,** 393-413.

[9] Cullen JB, Levitt SD. 1999 Crime, urban flight, and the consequences for cities. *Rev. Econ. Stat.* **81,** 159-169.

[10] Centers for Disease Control and Prevention. 2012 *HIV surveillance in urban and nonurban areas*; available at 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.* **104,** 7301-7306.

[12] Bettencourt LMA. 2013 The origin of scaling in cities. *Science* **340,** 1438-1441.

[13] Arbesman S, Kleinberg JM, Strogatz SH. 2009 Superlinear scaling for innovation in cities. *Phys. Rev. E* **79,** 016115.

[14] Pan W, Ghoshal G, Krumme C, Cebrian M, Pentland A. 2013 Urban characteristics attributable to density-driven tie formation. *Nat. Com.* **4,** 1961.

[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: Free Press.

[17] Topa G. 2001 Social interactions, local spillovers and unemployment. *Rev. Econ. Stud.* **68,** 261-295.

[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.

[19] Berk RA. 1983 An introduction to sample selection bias in sociological data. *Am. Sociol. Rev.* **48,** 386-398.

[20] Lazer D, et al. 2009 Computational social science. *Science* **323,** 721-723.

[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.* **104,** 7332-7336.

[22] Miritello G, Lara R, Cebrian M, Moro E. 2013 Limited communication capacity unveils strategies for human interaction. *Sci. Rep.* **3,** 1950.

[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.

[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(R).

[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.* **111,** 942-947.

[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.

[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.

[28] Geser H. 2006 Is the cell phone undermining the social order?: Understanding mobile technology from a sociological perspective. *Know. Techn. Pol.* **19,** 8-18.

[29] Eagle N, Pentland A, Lazer D. 2009 Inferring friendship structure by using mobile phone data. *Proc. Natl Acad. Sci.* **106,** 15274-15278.

[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.

[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.

[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 Mathematics* **1,** 226-251.

[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.* **102,** 4221-4224.

[35] Lee SH, Kim PJ, Jeong H. 2006 Statistical properties of sampled networks. *Phys. Rev. E* **73,** 016102.

[36] Watts DJ, Strogatz SH. 1998 Collective dynamics of 'small-world' networks. *Nature* **393,** 440-442.

[37] Raschke M, Schläpfer M, Nibali R. 2010 Measuring degree-degree association in networks. *Phys. Rev. E* **82,** 037102.

[38] Serrano MA, & Boguña M. 2005 Tuning clustering in random networks with arbitrary degree distributions. *Phys. Rev. E* **72,** 036133.

[39] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D. 2006 Complex networks: structure and dynamics. *Phys. Rep.* **424,** 175-308.

[40] Pastor-Satorras R, Vespignani A. 2001 Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86,** 3200-3203.

[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.

[42] Newman MEJ. 2009 Random graphs with clustering. *Phys. Rev. Lett.* **103,** 058701.

[43] Granovetter M. 1973 The strength of weak ties. *Am. J. Sociol.* **78,** 1360-1380.

[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.

[45] Granovetter M. 2005 The impact of social structure on economic outcomes. *J. Econ. Persp.* **19,** 33-50.

[46] Jacobs J. 1961 *The death and life of great American cities.* New York: 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.

[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.

ACKNOWLEDGMENTS

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 Program, Ericsson, BBVA, GE, Audi Volkswagen, Ferrovial and the members of the MIT Senseable City Lab Consortium. 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 Program (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.

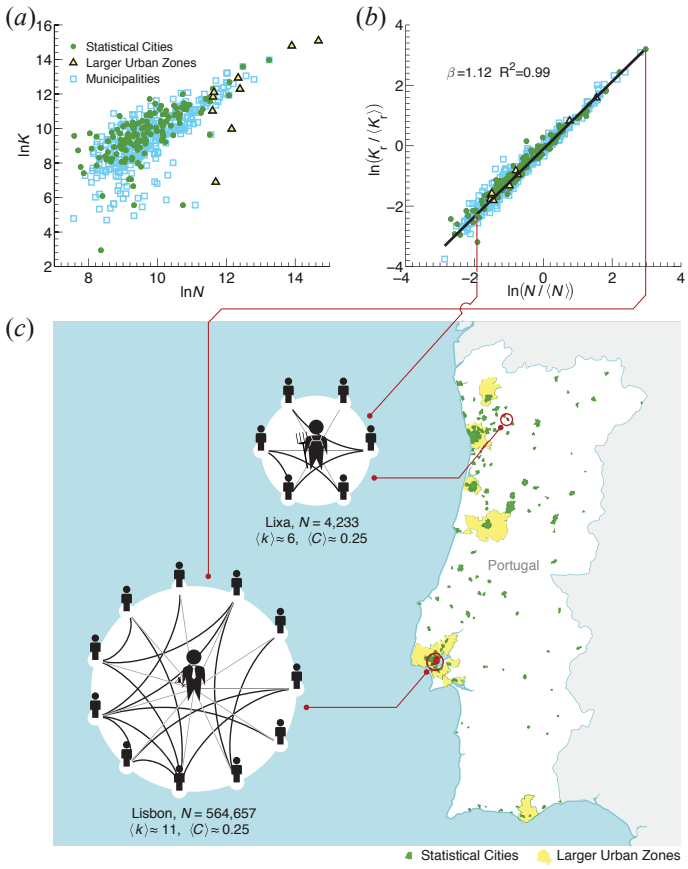

FIG. 1. Human interactions scale superlinearly with city size. (a) Cumulative degree, K, versus city population size, N, for three different city definitions in Portugal. (b) Collapse of the cumulative degree onto a single curve after rescaling by the coverage, $K_r = K/s$. For each city definition, the single values of K_r and N are normalised 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 Statistical Cities and Larger Urban Zones in Portugal, with the exception of those located on the archipelagos of the Azores and Madeira.

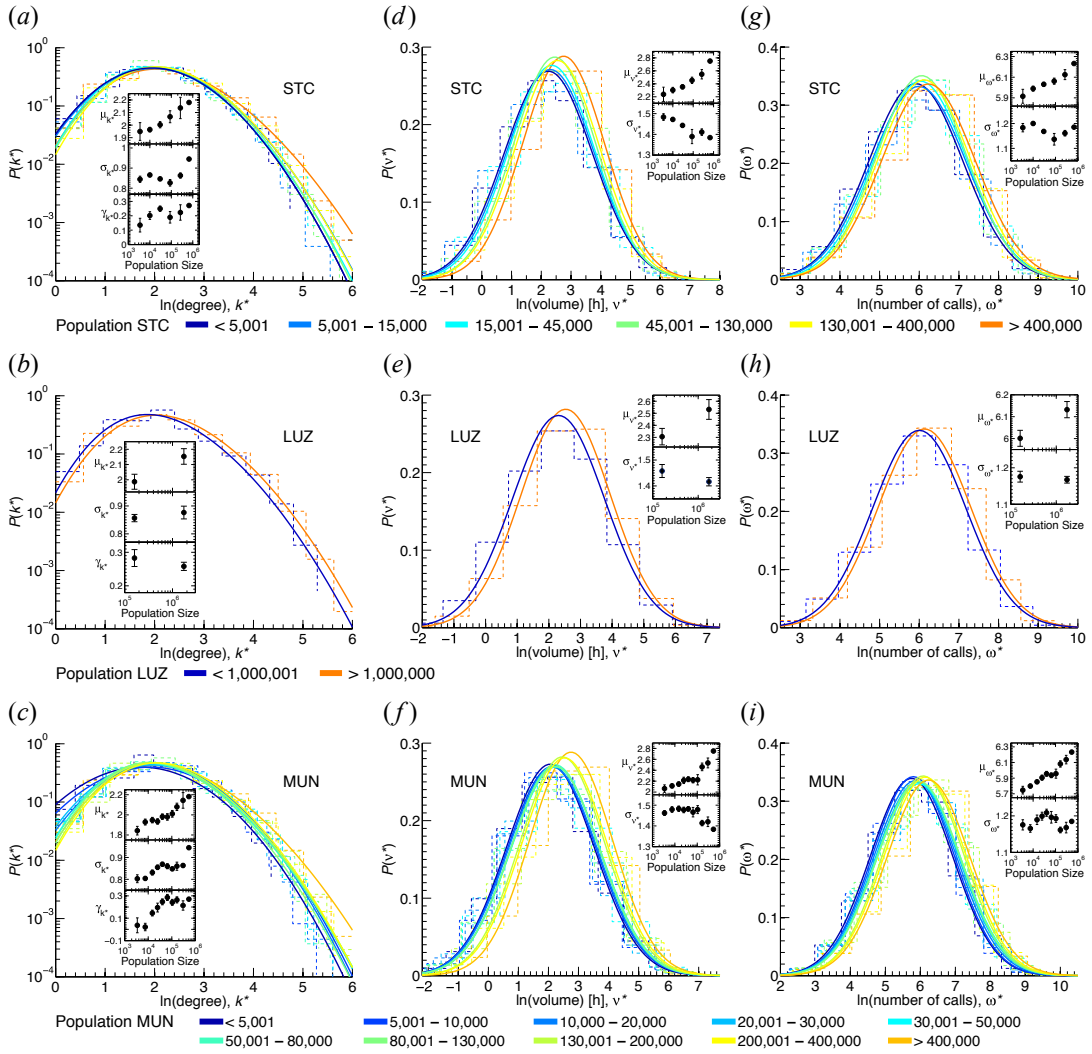

FIG. 2. The impact of city size on human interactions at the individual level. (*a-c*) Degree distributions, $P(k^*)$, for Statistical Cities (STC), Larger Urban Zones (LUZ) and Municipalities (MUN); the individual urban units are log-binned according to their population size. The dashed lines indicate 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 (SEM). The distribution parameters are estimated by the maximum likelihood method, see the electronic supplementary material.

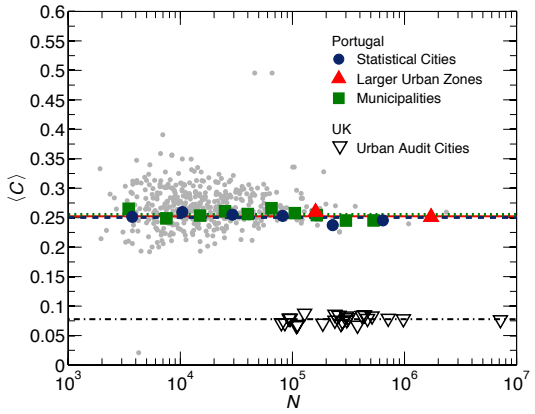

FIG. 3. The average clustering coefficient remains unaffected by city size. The lines indicate the average values with 0.251 ± 0.021 for STC (weighted average and standard deviation, dashed line), 0.252 ± 0.013 for LUZ (continuous line) and 0.255 ± 0.021 for MUN (dotted line) in Portugal, and 0.078 ± 0.004 for the UK (dash-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 (SEM) 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 value of $\langle C \rangle$ in the UK is lower than 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 3 people who communicate with a separate set of unconnected friends, we would indeed expect that the clustering coefficient would be approximately 1/3 of that of each individual.

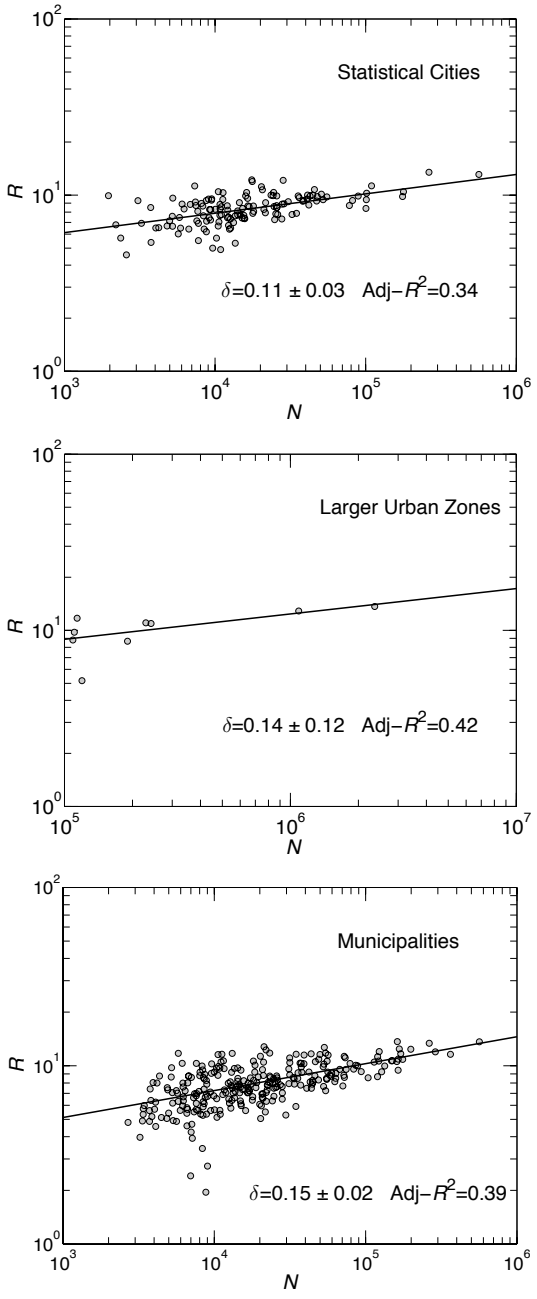

FIG. 4. Larger cities facilitate interaction-based spreading processes. The panels show the average spreading speed versus city size, broken down into the different city definitions. For each urban unit, the values of R result from averaging over 100 simulation trials performed on the reciprocal network 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% confidence intervals and the coefficients of determination are indicated.

TABLE I. Scaling exponents β. The observation period of $\Delta T = 409$ days is the full extent of the Portugal data set, while $\Delta T = 92$ days corresponds to the first three consecutive months. For the call volume statistics, we discarded 1 Larger Urban Zone (Ponta Delgada) due to a high estimation error of V_r (SEM $>$ 20%). For the UK data, the interaction indicators, Y, are not rescaled by the coverage due to consistently high market share. 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.

Portugal	City Definition	Number	Network Type	ΔT	Y	β	95% CI
	Statistical City	140	reciprocal	409 days	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 days	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 days	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 days	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 days	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 days	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	City Definition	Number	Network Type	ΔT	Y	β	95% CI
	Urban Audit City	24	reciprocal	31 days	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]

Electronic Supplementary Material

The scaling of human interactions with city size

Markus Schläpfer, Luís M. A. Bettencourt, Sébastian Grauwin, Mathias Raschke,
Rob Claxton, Zbigniew Smoreda, Geoffrey B. West and Carlo Ratti

(1) Supplementary Methods

Variation in the mobile phone coverage

Figure S1 shows the mobile phone coverage $s = |S|/N$ for each urban unit in Portugal (reciprocal (REC) network, overall observation period of ΔT = 409 days) with mean and standard deviation $\langle s \rangle = 0.18 \pm 0.13,\ 0.13 \pm 0.09,\ 0.14 \pm 0.11$ for Statistical Cities (STC), Larger Urban Zones (LUZ) and Municipalities (MUN), respectively. We find no significant correlations between the coverage and the population size (r = -0.02, p-value = 0.82 for STC, r = 0.34, p-value = 0.37 for LUZ, r = 0.09, p-value = 0.14 for MUN). The (non-significant) positive value of r for the 9 LUZ is mainly induced by a very low coverage of two smaller units located on the Azores and the island of Madeira (figure S1*e*). The otherwise low correlation levels indicate no asymmetric distribution of mobile phone users with respect to the size of the urban units.

We also do not find a clear trend consistent across all city definitions when applying linear regression to the log-transformed data, see figures S1*d* - S1*f*. Note that for the Municipalities there is a slight yet significant increase in s with population size. In this case, one could suspect that the superlinear scaling is the result of a larger number of subscribers in larger Municipalities. To test for this possibility, we increasingly filtered out a small number of Municipalities that have a lower coverage than a minimum threshold $s_{\min}$. Table S5 shows that (*i*) the positive relation between the population size and the coverage vanishes already for very small values of the minimum threshold ($s_{\min} \approx 5\%$), while (*ii*) the superlinear scaling of the interaction indicators persists. This shows that the observed increase in s is introduced by a small number of Municipalities with the lowest coverage, and that the superlinear scaling holds for the majority of Municipalities for which there is no positive relation between population size and coverage. Thus, this increase does not affect our conclusions reported in main text.

To further exclude the possibility that the superlinear scaling is the result of an increased number of subscribers in large cities compared to rural areas (e.g., due to better network accessibility), we analysed the scaling relation $Y \propto |S|^{\gamma}$ between the interaction

indicators, Y, and the number of callers, $|S|$, in each urban unit. If γ is equivalent to the exponent β for the population size N (see main text), a potential effect of the varying number of callers with city size can be further excluded, while s can be interpreted as a random variable that is independent of N. Comparing table S6 with table 1 in the main text confirms the excellent agreement of the two scaling exponents.

As a consequence, we would expect similar power-law exponents when the data is *not* rescaled by the coverage. However, as mentioned in the main text, a power law is difficult to justify in that case which is reflected in substantially lower coefficients of determination. For instance, fitting the relation between the 'non-normalised' cumulative degree, K, (figure 1a) and city size in Portugal (Statistical Cities) by a power law leads to $\beta = 1.17$ (95% CI [1.10, 1.23]) with Adj-R^2 = 0.90 (in comparison to Adj-$R^2 \approx 1$ after rescaling).

Moreover, superlinear power-law scaling is also valid when considering only cities with a high coverage. As an example, limiting the analysis to Statistical Cities with $s > 0.30$, which holds for 19 urban units and implies an average coverage of $\langle s \rangle \approx 0.42$ (compared to $\langle s \rangle \approx 0.20$ for all Statistical Cities), we get $\beta = 1.17$ (95% CI [1.10, 1.25]) for the rescaled cumulative degree (REC network, $\Delta T = 409$ days). Similarly, for the non-reciprocal (nREC) network, which contains a larger number of nodes than its reciprocal counterpart, we get $\beta = 1.18$ (95% CI [1.08, 1.28]) for $\langle s \rangle \approx 0.50$, corresponding to the 15 best sampled (i.e., with the highest coverage) Statistical Cities. Moreover, $\beta > 1$ holds for even higher values of the average coverage, but the small number of urban units that fulfil this condition strongly limits here our conclusions. Nevertheless, together with the results from the UK (>95% market share), these findings indicate that superlinear scaling is largely robust against different subsamples (i.e., different market shares) of the complete interaction network.

To test whether the number of mobile phone subscribers depends on socioeconomic urban characteristics such as wages, we utilised publicly available income data at the Municipality level in Portugal for 2009 [1]. We find no correlation between the coverage

(REC network, $\Delta T = 409$ days) and the per-capita monthly income (r = 0.01, p-value = 0.82), again supporting the assumption of a symmetric composition of the user base with respect to the different urban units. In contrast, we find a significant correlation between the average degree $\langle k \rangle$ and the per-capita monthly income (r = 0.35, p-value < 10^{-4}), supporting the hypothesised correspondence between the average social connectivity and socioeconomic urban quantities.

Finally, the superlinear scaling relations also hold when restricting k_i to calling partners within the same city. For instance, we find for the rescaled cumulative degree of the Statistical Cities (REC network, $\Delta T = 409$ days) a scaling exponent of $\beta = 1.26$ (95% CI [1.23, 1.30]). Nevertheless, as smaller cities may hereby induce trivial boundary effects that lead to an overestimation of β, we included all links for the results reported in the main text.

Individual-based interaction distributions

The individual-based interaction indicators are inherently time-aggregated values that become affected by longer periods of call inactivity due to, e.g., cancelling an ongoing subscription during the observation window ΔT. Those callers that are not active on a regular basis naturally induce a bias resulting in a skewness in the distributions of k, v and ω, as their accumulated measures remain at lower values (figure S4). To compare the distributions in a meaningful way [2], we focus here on regularly active callers by estimating the probability distributions based on those individuals that initiate and receive at least one call every $f_{\min}$ subsequent days. Less active individuals are included in terms of their connections to those regularly active callers. We chose $f_{\min}$ = 1/[90 days] which substantially decreases the skewness, while considering over 50% of all nodes in the reciprocal network (i.e., $n_f = 8.7 \times 10^5$ nodes) as regularly active. The superlinear scaling of the mean is not affected by $f_{\min}$. In addition, we tested alternative methods of homogenising the set of callers. For instance, we selected only individuals that appeared both in the first and last month of the overall observation period, yielding again qualitatively similar results to those reported in the main text. In all

cases, while the exact shape of the distributions generally depends on the network sampling [3], the mean of the distributions showed superlinear scaling compatible with table 1 in the main text.

To choose the probability model that best describes the homogenised distributions we selected as trial models (*i*) the lognormal distribution, (*ii*) the generalised Pareto distribution, (*iii*) the double Pareto-lognormal distribution and (*iv*) the log-skew-normal distribution (or 'skewed lognormal distribution'). The lognormal distribution (LN) of a random variable X implies that $Y = \ln X$ is normally distributed with probability density

$$P(y) = \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right). \tag{S1}$$

Lognormal distributions are naturally generated by multiplicative random processes and thus are widespread in sociology and economics [4]. The generalised Pareto (GP) distribution includes the exponential and the Pareto distribution as special cases [5]. The latter is a commonly used power-law distribution. The double Pareto-lognormal distribution (DPLN) has recently been shown to accurately model the empirical distributions of the degree, call volume and number of calls in a mobile phone network [6]. Finally, X follows a log-skew-normal distribution (LSN) [7] if $Y = \ln X$ obeys a skew-normal distribution

$$P(y) = \frac{2}{\theta}\phi\left(\frac{y-\xi}{\theta}\right)\Phi\left(\alpha\left(\frac{y-\xi}{\theta}\right)\right), \tag{S2}$$

where ξ, θ and α are the location, scale and shape parameters, respectively. To simplify the interpretation the 'direct parameters' (ξ, θ, α) can be transformed into the 'centred parameters' (μ, σ, γ) where $\mu = \mathrm{E}\{Y\}$, $\sigma^2 = \mathrm{var}\{Y\}$ and γ denotes the skewness of the distribution [8]. By allowing for non-zero skewness, equation (S2) constitutes a generalisation of the normal distribution.

Tables S7-S9 indicate how many times each distribution has outperformed all other models in terms of the log-likelihood function and the BIC (REC network, $\Delta T = 409$ days, $f_{\min} =$ 1/[90 days]). The two Larger Urban Zones located on the archipelagos (Ponta Delgada and Funchal) are not considered due to a substantially lower market share ($s < 0.01$ for the homogeneous set of callers). For the call volume, we further discarded 1 Statistical City to which only 4 regularly active callers were assigned. The log-skew-normal distribution is in most cases the best model for the degree distribution, as there remains a slight right-skewness even when considering only regularly active callers. In particular, equation (S2) provides an excellent fit for the right tail of the distribution (figure S5). Generally, right-skewness can be explained by a 'hidden' constraint on small values (or lower truncation) of otherwise normally distributed observations [9]; we intend to elaborate on this point in future work. The BIC favours the lognormal distribution for both the call volume and number of calls.

Clustering coefficient – comparison to the random case and regression analysis

In an Erdős-Rényi random network with the same number of nodes $|S| = sN$ and same average degree $\langle k \rangle$ as in the studied cities, the expected clustering coefficient is $\langle C \rangle_{\mathrm{ER}} = \langle k \rangle / |S|$ [10]. Given the superlinear scaling relation we observed for the number of contacts, the value of $\langle C \rangle_{\mathrm{ER}}$ can be expected to vary with city population size as

$$\langle C \rangle_{\mathrm{ER}} = \frac{\langle k \rangle}{|S|} \sim \frac{N^{\beta-1}}{N} \sim N^{\gamma_{\mathrm{ER}}}, \gamma_{\mathrm{ER}} = \beta - 2 . \qquad \text{(S3)}$$

With $\beta \approx 1.12$ (see table 1 in the main text) the expected value of the exponent is $\gamma_{\mathrm{ER}} \approx -0.88$. Thus, the average clustering coefficient in the corresponding Erdős-Rényi network would decrease rapidly with increasing city size. For a comparison with the real data, we performed a regression analysis on the studied communication networks. In contrast to the Erdős-Rényi random network, we find that the clustering coefficient remains largely unaffected by city size, with Adj-R^2 values of the regression analysis being very low and values of γ being very close to 0 for all different city definitions, see figure S7.

Weighted clustering coefficients

The standard clustering coefficient does not consider the weights of the links in terms of the accumulated call volume and number of calls between two individuals. Hence, to assess the influence of the weights, we also computed for each caller i the weighted clustering coefficient according to ref. [11],

$$\tilde{C}_i = \frac{1}{k_i(k_i-1)} \frac{1}{\max(w)} \sum_{jk} (w_{ij} w_{ik} w_{jk})^{1/3}, \tag{S4}$$

where the weight w_{ij} is either the accumulated number of calls or the accumulated call duration between callers i and j. This weighted clustering coefficient is a natural generalisation of the standard unweighted coefficient (notice that in the simple case $w = 1$, $\tilde{C} = C$). We find that the weighted clustering coefficients for both the number of calls and the call duration, averaged over all callers in a given city, are largely invariant of city size in both Portugal and UK, see figure S8 and table S10, which confirms the behaviour of the standard clustering coefficient. Moreover, in case of Portugal's mobile phone network, the average values do not strongly depend on the particular city definition.

Degree-degree correlations

To quantify degree-degree correlations in the analysed networks, we computed the average degree $\langle k_{nn}(k) \rangle$ of the nearest-neighbours of nodes having degree k, which is one of the most widely used measures, see ref. [10]. If $\langle k_{nn}(k) \rangle$ is an increasing function of k, the nodes tend to be connected to other nodes with similar degree, corresponding to assortative (or positive) degree-degree correlations. As depicted in figure S9, Portugal's mobile phone network indeed exhibits assortative degree-degree correlations, with $\ln \langle k_{nn}(k) \rangle \sim \ln(k)$ being valid for a wide range of values of k (the linear regression has an Adj-R^2 = 0.99 for k <100, which accounts for 99.8% of the nodes). This relation indicates a strong tendency of a node to connect to other nodes with degree of similar magnitude. In contrast, the landline network in the UK does not exhibit such a clear positive correlation between k and $\langle k_{nn}(k) \rangle$.

(2) Supplementary Figures

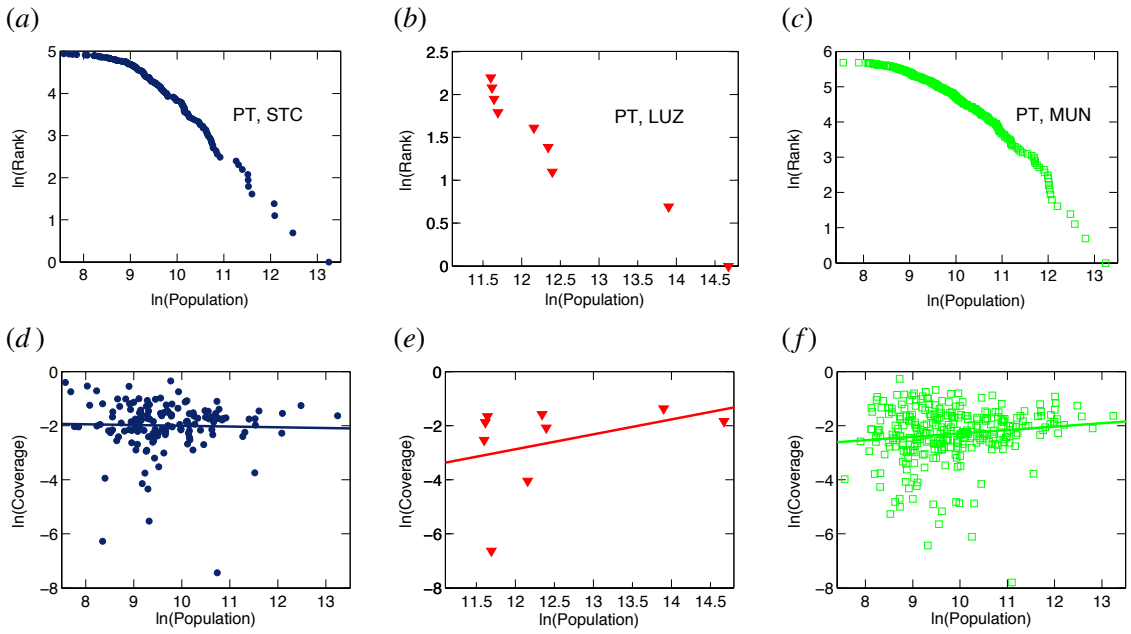

Figure S1: Population size distribution for the urban units in Portugal and relative number of assigned callers. (*a-c*) Zipf plots for Statistical Cities (*a*), Larger Urban Zones (*b*) and Municipalities (*c*). (*d-f*) Corresponding mobile phone coverage resulting from the node assignment procedure (REC network, ΔT = 409 days). The solid lines show the linear regressions with slopes -0.03 $\pm$ 0.16 (95% CI, Adj-R^2 = -0.01) for the Statistical Cities (*d*), 0.55 $\pm$ 1.28 (95% CI, Adj-R^2 = 0.004) for the Larger Urban Zones (*e*) and 0.13 $\pm$ 0.11 (95% CI, Adj-R^2 = 0.01) for the Municipalities (*f*).

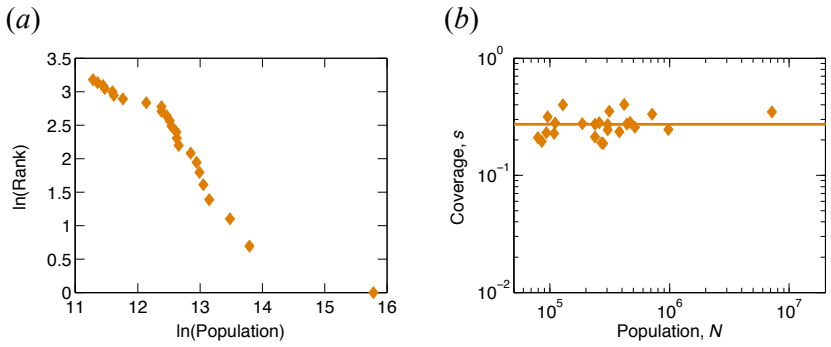

Figure S2. City size distribution for the UK and relative number of landline phones. (*a*) Zipf (rank-size) plot for the population of the Urban Audit Cities. (*b*) Corresponding landline phone coverage. The solid line corresponds to the average value.

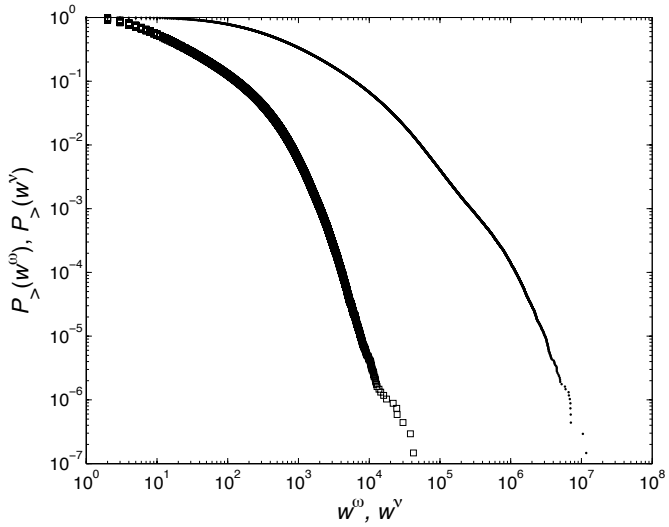

Figure S3. Cumulative distributions of the tie strength (link weights) in terms of the accumulated number of calls w^ω (squares), and accumulated call volume w^ν in seconds (circles) between each pair of callers, for the REC network in Portugal with $\Delta T = 409$ days.

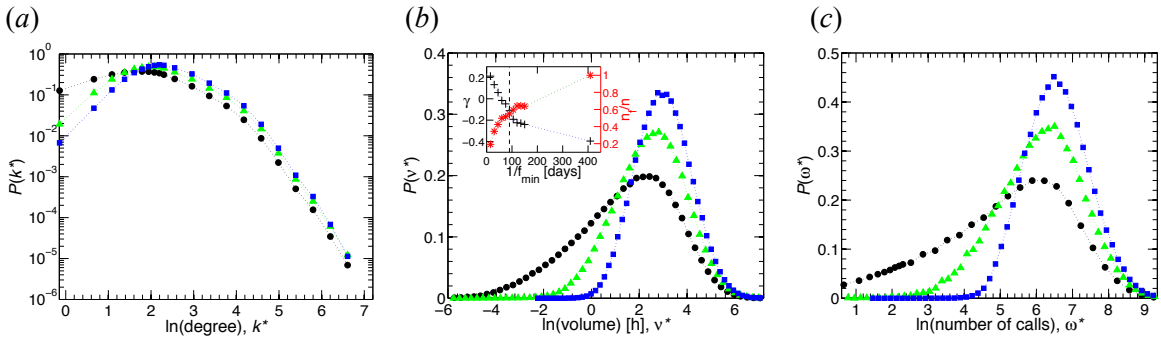

Figure S4. Increasing the homogeneity of the interaction distributions. (*a*) Degree distribution for the REC network in Portugal (ΔT = 409 days). To highlight the tail behaviour we show the probabilities on a logarithmic scale. (*b,c*) corresponding distribution of the call volume and number of calls. When considering all callers (black circles) the distributions are strongly left-skewed. Considering only callers whose call frequency is higher than $f_{\min}$ = 1/[90 days] (green triangles) and $f_{\min}$ = 1/[30 days] (blue squares) gradually decreases the skewness. Most notably for $v^* = \ln v$ and $\omega^* = \ln \omega$, the homogenised data increasingly resemble the Gaussian bell curve (i.e., a lognormal distribution in the original variables). The inset in (*b*) depicts the decrease of the average skewness (third standardised moment), γ, for all Statistical Cities with increasing $f_{\min}$, and the corresponding fraction of regularly active callers, n_f / n.

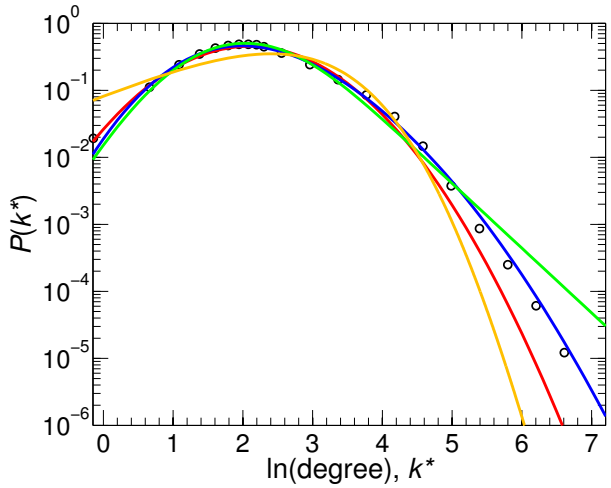

Figure S5. Degree distribution. Black circles: distribution of the regularly active callers in Portugal (REC network, ΔT = 409 days, $f_{\min}$ = 1/[90 days]). The continuous lines are best fits of the lognormal (red), generalised Pareto (yellow), double Pareto-lognormal (green) and log-skew-normal model (blue).

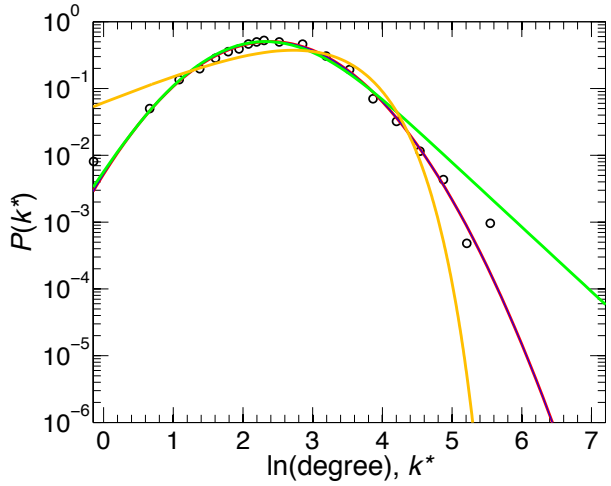

Figure S6. Degree distribution for the city with the highest coverage. Black circles: distribution of the regularly active callers (REC network, $\Delta T = 409$ days, $f_{\min}$ = 1/[90 days]). The Statistical City has $N = 17{,}535$ inhabitants and a total of $|S|$ = 12,304 assigned callers, resulting in $s = 0.70$. The continuous lines are as in figure S5. The best fits of the lognormal and log-skew-normal model coincide and outperform the other distributions, in agreement with the behaviour for the average coverage of $\langle s \rangle \approx 0.20$ (table S7).

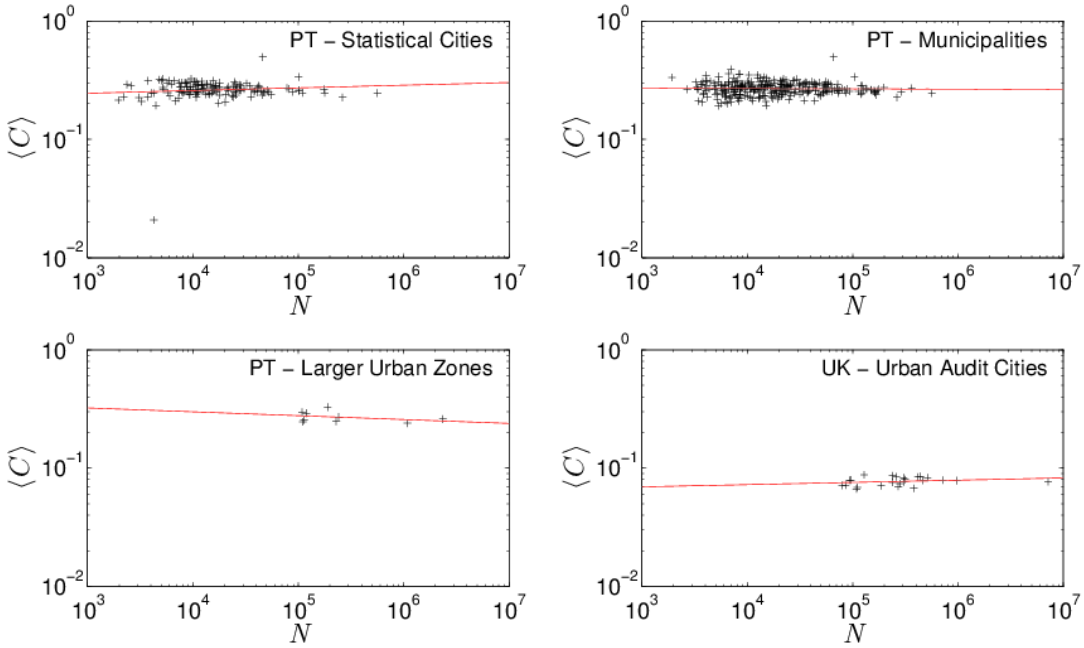

Figure S7. Regression analysis on the average clustering coefficients. Black crosses are the values of $\langle C \rangle$ versus city size for the different city definition as used in the main text. Red lines show the best linear regression, $\ln\langle C \rangle = \alpha + \gamma \ln N$. The values for the slopes and the corresponding 95% confidence interval are $\gamma = 0.023$ [0.018, 0.027] for the Statistical Cities, $\gamma = -0.002$ [-0.004, 0.001] for the Municipalities and $\gamma = -0.033$ [-0.039, -0.027] for the Larger Urban Zones in Portugal (REC network, $\Delta T = 409$ days), and 0.019 [0.016, 0.021] for the Urban Audit Cities in the UK. For all fits Adj-$R^2 < 0.1$.

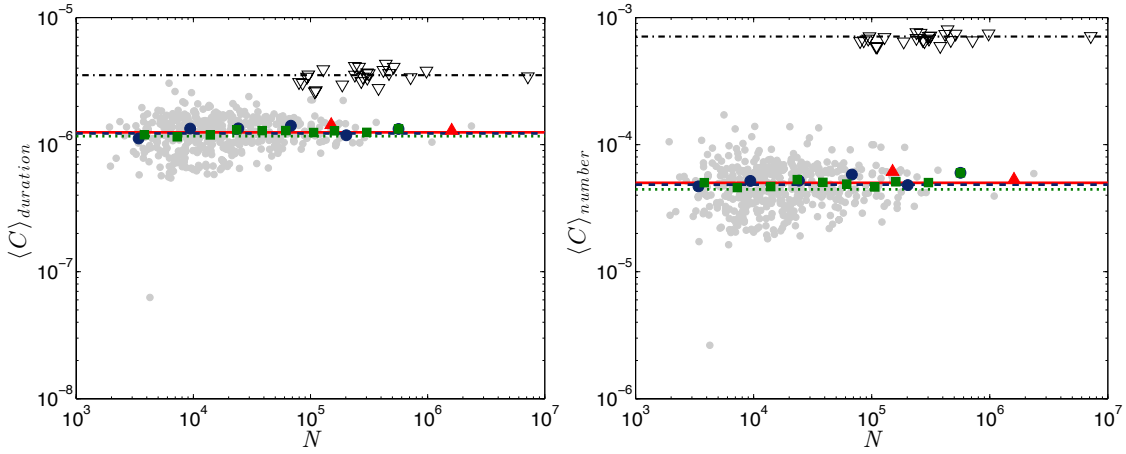

Figure S8. Average weighted clustering coefficients based on the call duration, $\langle C \rangle_{\text{duration}}$ (left), and based on the number of calls, $\langle C \rangle_{\text{number}}$ (right), for Portugal (REC network, $\Delta T = 409$ days) and UK, with symbols according to figure 3 in the main text. The lines correspond to the averages of the different city definitions, values are reported in table S10. Grey points are the underlying scatter plot for all urban units.

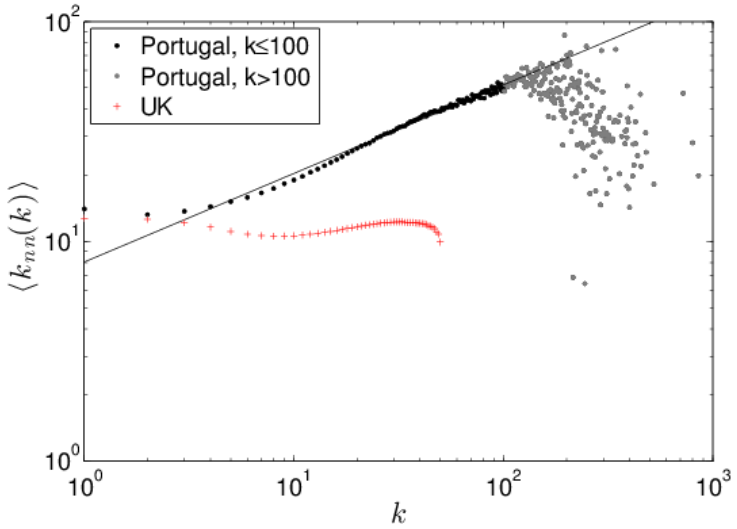

Figure S9. Average nearest-neighbour degree as a function of the nodal degree, $\langle k_{nn}(k) \rangle$. For Portugal (REC network, $\Delta T = 409$ days), the value varies like $\ln \langle k_{nn}(k) \rangle = \alpha + \gamma \ln k$ ($\gamma \approx 0.4$, Adj-$R^2 = 0.99$) for low values of k ($k < 100$, accounting for 99.8% of all nodes).

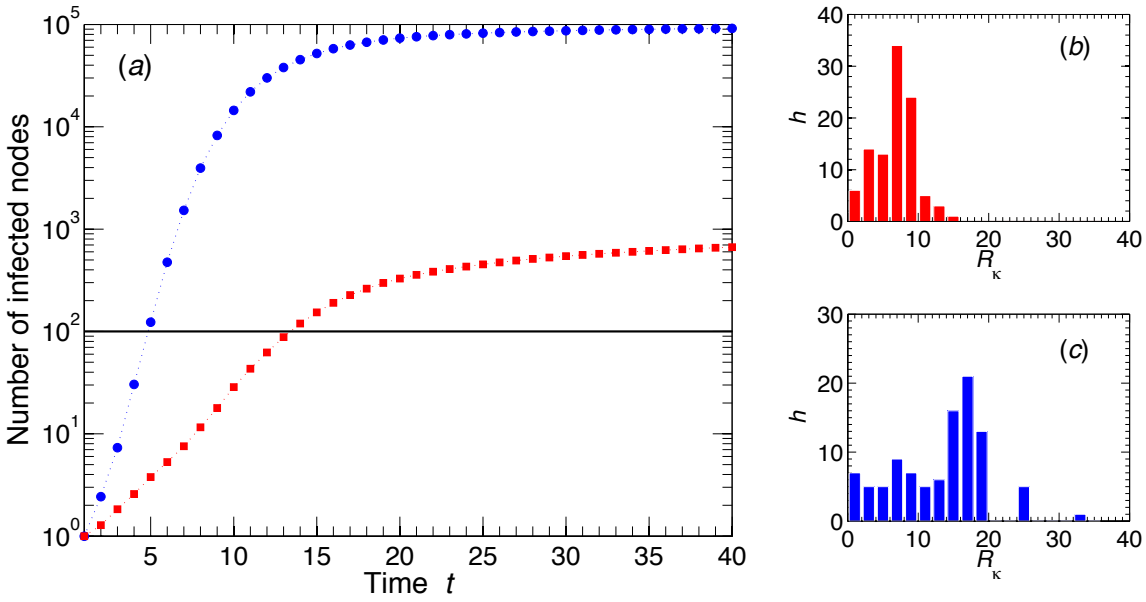

Figure S10. (*a*) The number of infected nodes as a function of the simulation time t for the examples of Lisbon (STC, with N=564,657 and $|S|$=109,448, blue circles), and Meda (STC, with N=2193, $|S|$=1033), averaged over 100 simulation runs (REC network, $\Delta T = 409$ days). Both urban units lie very close to the regression line in figure 4*a* of the main text. The spreading is substantially faster in Lisbon. The continuous line indicates the stopping criterion for estimating R (i.e., when $n_I = 100$ nodes were infected). (*b*) Histogram of the single values of the spreading speed, $R_\kappa = n_I / t_\kappa(n_I)$ (see main text), resulting from each of the 100 individual simulation runs for Meda and (*c*) for Lisbon.

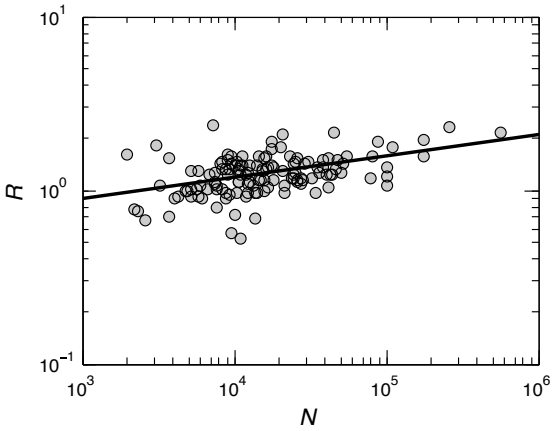

Figure S11. Spreading speed R for the Statistical Cities in Portugal based on the unweighted reciprocal network (REC network, $\Delta T = 409$ days). The spreading model has been implemented as for the weighted case (see main text), with the only difference that instead of the weight-dependent transmission probability, we used a fixed value $P_{ij} = 0.01$. The solid line is the best fit to a power law scaling relation $R \propto N^{\delta}$ with exponent $\delta = 0.12 \pm 0.04$ (95% CI, Adj-R^2 = 0.22).

(3) Supplementary Tables

Network type	ΔT [days]	n	m	$\langle k \rangle$	$\langle v \rangle$ [hours]	$\langle \omega \rangle$	LCC
REC	409	1,589,511	6,770,405	8.52	12.03	498.56	0.98
	92	1,087,722	2,867,400	5.27	5.54	158.03	0.93
nREC	409	1,802,802	11,354,604	12.60	17.22	473.85	0.99

Table S1. Summary statistics for the mobile phone networks in Portugal. The size of the largest connected component (LCC) is given as a fraction of the total number of nodes. All networks are considered as being undirected, so that the LCC for the nREC network corresponds to the giant weakly connected component (GWCC). The values for the number of nodes n, number of links m, average degree $\langle k \rangle$, average call volume $\langle v \rangle$ and average number of calls $\langle \omega \rangle$ correspond to those of the LCC. The distribution of the tie strengths is shown in figure S3.

City definition	No. of entities	N^{tot}	N^{min}	N^{max}
Statistical Cities	140	4,032,176	1,960	564,657
Municipalities	293	9,901,216	1,924	564,657
Larger Urban Zones	9	4,566,630	108,891	2,363,470

Table S2. Population statistics of the analysed urban units in Portugal for the year 2001. For each city definition we show the total population covered, N^{tot}, as well as the population size of the smallest (N^{min}) and largest (N^{max}) entity.

Network type	n^{tot}	m^{tot}	n^{land}	$\langle k^{\text{land}} \rangle$	$\langle v^{\text{land}} \rangle$ [hours]	$\langle \omega^{\text{land}} \rangle$	LCC
REC	47,072,811	119,725,827	24,054,946	7.97	6.61	102.1	0.99

Table S3. Summary statistics of the UK communication network. The number of nodes (n^{tot}) and number of links (m^{tot}) correspond to the LCC of the overall network (including mobile phones connected to landlines). All other values correspond to the landlines only (including their links to mobile phones). The network is undirected.

City definition	No. of entities	N^{tot}	N^{min}	N^{max}
Urban Audit Cities	24	14,186,179	79,734	7,172,091

Table S4. Population statistics of the analysed urban system in the UK for the year 2001. The variables are defined as in table S2.

s_{min}	Number of Municipalities with $s > s_{min}$	γ	β		
			K_r	V_r	W_r
0	293	0.13 [0.02, 0.24]	1.13 [1.11, 1.14]	1.15 [1.13, 1.17]	1.13 [1.11, 1.14]
0.01	279	0.10 [0.02, 0.19]	1.13 [1.11, 1.15]	1.16 [1.14, 1.18]	1.13 [1.11, 1.14]
0.02	271	0.06 [-0.02, 0.14]	1.13 [1.11, 1.15]	1.16 [1.14, 1.18]	1.13 [1.11, 1.14]
0.03	265	0.05 [-0.03, 0.12]	1.12 [1.11, 1.14]	1.16 [1.14, 1.18]	1.13 [1.11, 1.15]
0.04	260	0.03 [-0.04, 0.10]	1.13 [1.11, 1.14]	1.16 [1.14, 1.18]	1.13 [1.11, 1.15]
0.05	251	0.02 [-0.05, 0.09]	1.12 [1.11, 1.14]	1.16 [1.14, 1.18]	1.13 [1.11, 1.15]
0.06	236	-0.01 [-0.07, 0.06]	1.12 [1.10, 1.14]	1.17 [1.14, 1.19]	1.13 [1.11, 1.15]

Table S5. Effect of filtering out Municipalities with a coverage lower than s_{min} on the slope of the linear regression $\ln(s) = \alpha + \gamma \ln(N)$ and on the scaling exponent β for the interaction indicators (K_r, V_r, W_r) according to table 1 in the main text (REC network, $\Delta T = 409$ days). The number in the square brackets indicate the 95% confidence interval. The slope γ of the best linear fit systematically decreases with increasing value of s_{min}, indicating that the dependence of the coverage on the population size vanishes for the majority of Municipalities that have higher values of s. In contrast, the superlinear scaling of the interaction indicators remains largely unaffected. This result shows that the positive relation between population size and coverage is introduced by a small number of Municipalities with lowest coverage. It thus further excludes the possibility that the superlinear scaling as reported in the main text is the result of an increasing number of subscribers in larger Municipalities.

Caller network	Y	γ	95% CI
reciprocal	Degree	1.10	[1.09, 1.11]
	Call volume	1.13	[1.11, 1.14]
	Number of calls	1.10	[1.09, 1.12]
non-reciprocal	Degree	1.20	[1.18, 1.22]
	Call volume	1.15	[1.13, 1.17]
	Number of calls	1.13	[1.11, 1.14]

Table S6. Resulting exponents of the scaling relations based on the number of callers. The values are shown for the Statistical Cities in Portugal ($\Delta T = 409$ days). Exponents were estimated by nonlinear least squares (trust-region algorithm).

			Distribution model			
City definition	No. of entities	Statistical method	LN	GP	DPLN	LSN
Statistical Cities	140	ln L	0	0	52	88
		BIC	50	1	20	69
Larger Urban Zones	7	ln L	0	0	5	2
		BIC	0	0	3	4
Municipalities	293	ln L	1	1	116	175
		BIC	142	5	15	131
All types	440	ln L	1	1	173	265
		BIC	192	6	38	204

Table S7. Model selection for the degree distribution by the 'goodness of the fit' (REC network, $\Delta T = 409$ days, $f_{\min} = 1/[90 \text{ days}]$). The numbers indicate how many times each distribution has been selected based on the maximum value of the log-likelihood function ($\ln L$) and the BIC, respectively.

City definition	No. of entities	Statistical method	Distribution model			
			LN	GP	DPLN	LSN
Statistical Cities	139	ln L	0	0	32	107
		BIC	91	7	6	35
Larger Urban Zones	7	ln L	0	0	1	6
		BIC	2	0	0	5
Municipalities	293	ln L	0	0	86	207
		BIC	225	13	3	52
All types	439	ln L	0	0	119	320
		BIC	318	20	9	92

Table S8. Model selection for the distribution of the call volume.

City definition	No. of entities	Statistical method	Distribution model			
			LN	GP	DPLN	LSN
Statistical Cities	140	ln L	0	0	29	111
		BIC	53	4	8	75
Larger Urban Zones	7	ln L	0	0	0	7
		BIC	0	0	0	7
Municipalities	293	ln L	0	2	89	202
		BIC	170	13	6	104
All types	440	ln L	0	2	118	320
		BIC	223	17	14	186

Table S9. Model selection for the distribution of the number of calls.

City definition	$\langle C \rangle_{\text{number}}$	$\langle C \rangle_{\text{duration}}$
PT - Statistical Cities	$(4.8\pm1.3)\times10^{-5}$	$(1.23\pm0.24)\times10^{-6}$
PT - Municipalities	$(4.5\pm1.4)\times10^{-5}$	$(1.17\pm0.27)\times10^{-6}$
PT – Larger Urban Zones	$(5.0\pm1.1)\times10^{-5}$	$(1.25\pm0.19)\times10^{-6}$
UK - Urban Audit Cities	$(7.1\pm0.4)\times10^{-4}$	$(3.52\pm0.29)\times10^{-6}$

Table S10. Weighted averages and standard deviations of the weighted clustering coefficients (see figure S8) for the different city definitions.

References

1. Pordata – Base de Dados Portugal Contemporâneo (2013); available at http://www.pordata.pt.
2. 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.
3. Lee SH, Kim PJ, Jeong H. 2006 Statistical properties of sampled networks. *Phys. Rev. E* **73,** 016102.
4. Mitzenmacher M. 2004 A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* **1,** 226-251.
5. Hüsler J, Li D, Raschke M. 2011 Estimation for the generalized Pareto distribution using maximum likelihood and goodness of fit. *Commun. Stat. – Theor. M.* **40,** 2500-2510.
6. Seshadri M, Machiraju S, Sridharan A, Bolot J, Faloutsos C, Leskovec J. (2008) Mobile call graphs: beyond power-law and lognormal distributions. *In Proc. of ACM SIGKDD*.
7. Azzalini A, Capitano A. 1999 Statistical applications of the multivariate skew-normal distribution. *J. R. Statist. Soc. B* **61,** 579-602.
8. Arellano-Valle RB, Azzalini A. 2008 The centred parametrization for the multivariate skew-normal distribution. *J. Multivariate Anal.* **99,** 1362-1382.
9. Azzalini A. 2005 The skew-normal distribution and related multivariate families. *Scand. J. Statist.* **32,** 159-188.
10. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D. 2006 Complex networks: structure and dynamics. *Phys. Rep.* **424**, 175-308.
11. Onnela J, Saramäki J, Hyvönen J, Szabò G, De Menezes M, Kaski K, Barabàsi, A-L. 2007 Analysis of a large-scale weighted network of one-to-one human communication. *New J. Phys.* **9,** 179.