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 ± 0.09, 0.14 ± 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 $\langle C \rangle_{\text{ER}} = \langle k \rangle / |S|$ [10]. Given the superlinear scaling relation we observed for the number of contacts, the value of $\langle C \rangle_{\text{ER}}$ can be expected to vary with city population size as

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

With $\beta \approx 1.12$ (see table 1 in the main text) the expected value of the exponent is $\gamma_{\text{ER}} \approx -0.88$. 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],

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

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

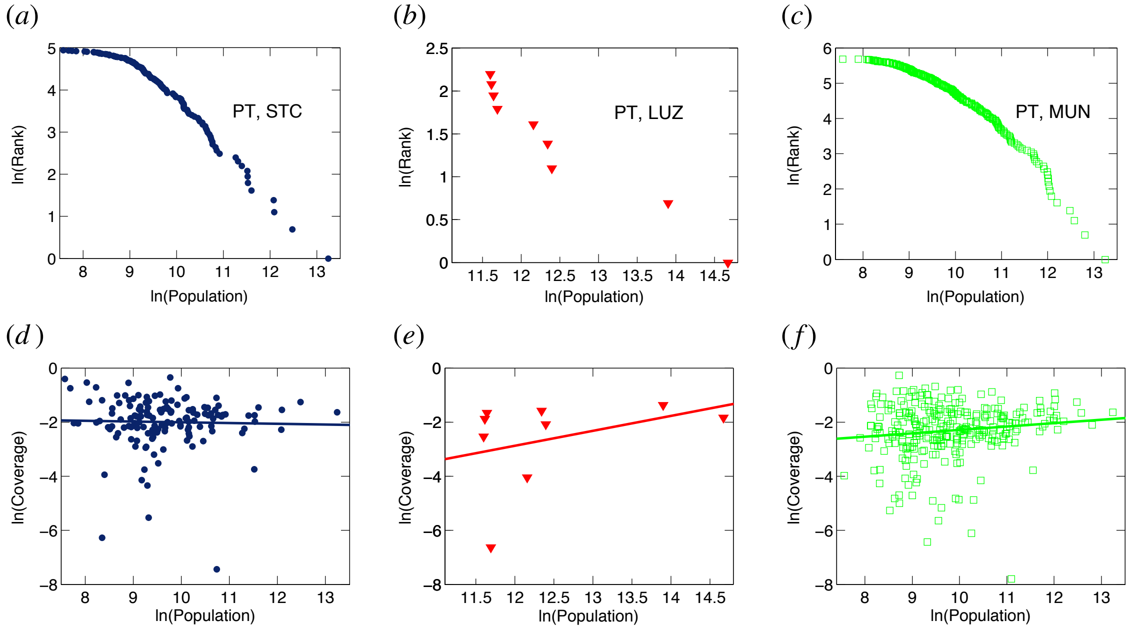

Figure S1: Population size distribution for the urban units in Portugal and relative number of assigned callers. (*a-c*) Zipf plots for Statistical Cities (*a*), Larger Urban Zones (*b*) and Municipalities (*c*). (*d-f*) Corresponding mobile phone coverage resulting from the node assignment procedure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

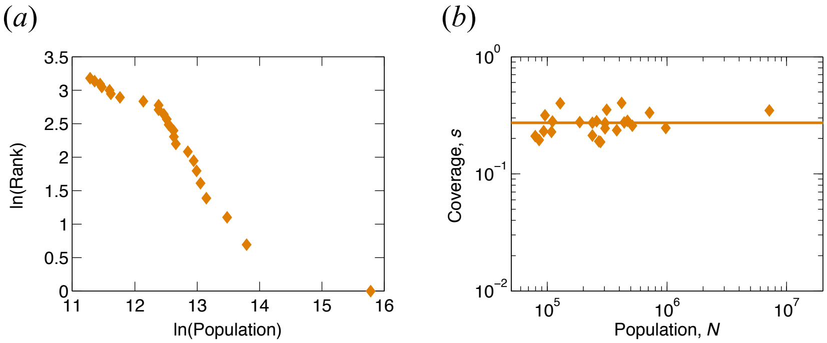

Figure S2. City size distribution for the UK and relative number of landline phones. (*a*) Zipf (rank-size) plot for the population of the Urban Audit Cities. (*b*) Corresponding landline phone coverage. The solid line corresponds to the average value.

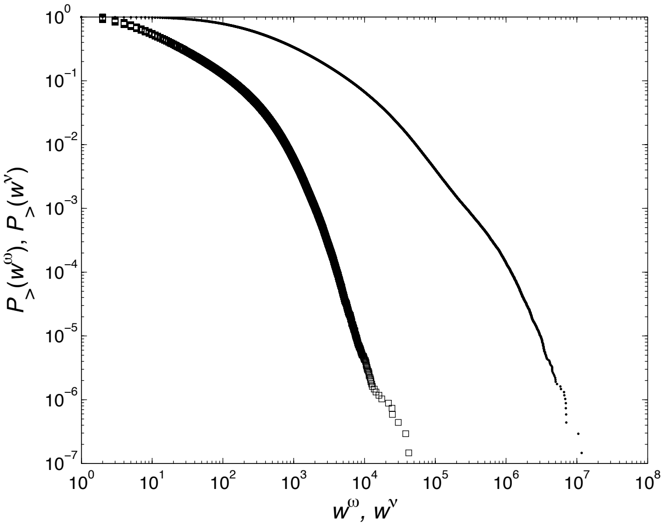

Figure S3. Cumulative distributions of the tie strength (link weights) in terms of the accumulated number of calls w^{ω} (squares), and accumulated call volume w^{ν} in seconds (circles) between each pair of callers, for the REC network in Portugal with $\Delta T = 409$ days.

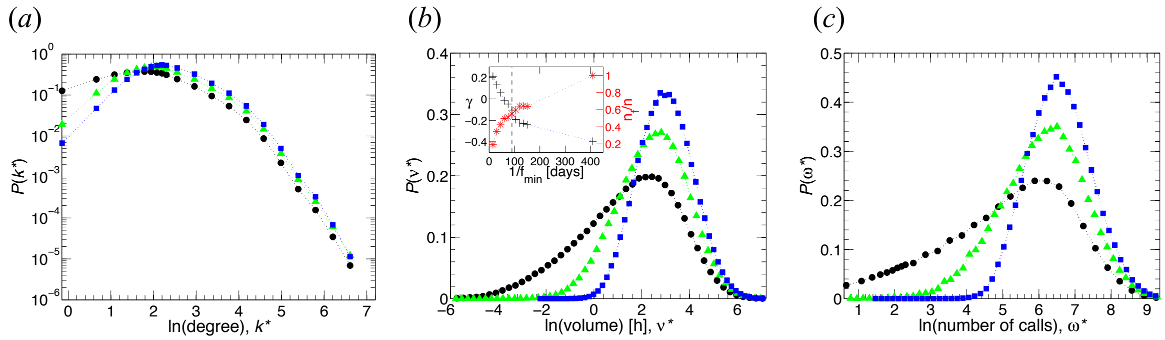

Figure S4. Increasing the homogeneity of the interaction distributions. (*a*) Degree distribution for the REC network in Portugal (ΔT = 409 days). To highlight the tail behaviour we show the probabilities on a logarithmic scale. (*b,c*) corresponding distribution of the call volume and number of calls. 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.

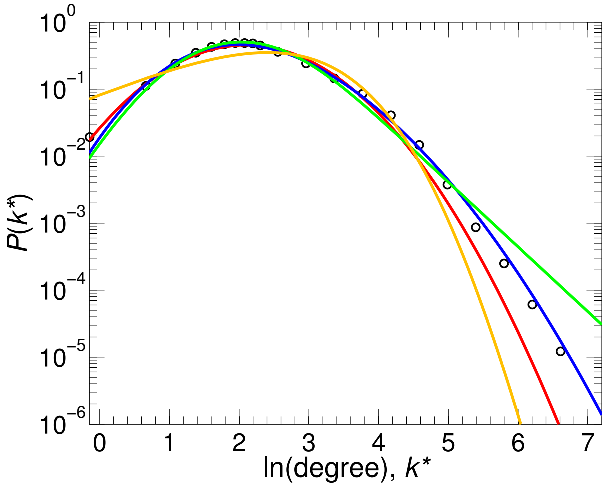

Figure S5. Degree distribution. Black circles: distribution of the regularly active callers in Portugal (REC network, ΔT = 409 days, $f_{\min}$ = 1/[90 days]). The continuous lines are best fits of the lognormal (red), generalised Pareto (yellow), double Pareto-lognormal (green) and log-skew-normal model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

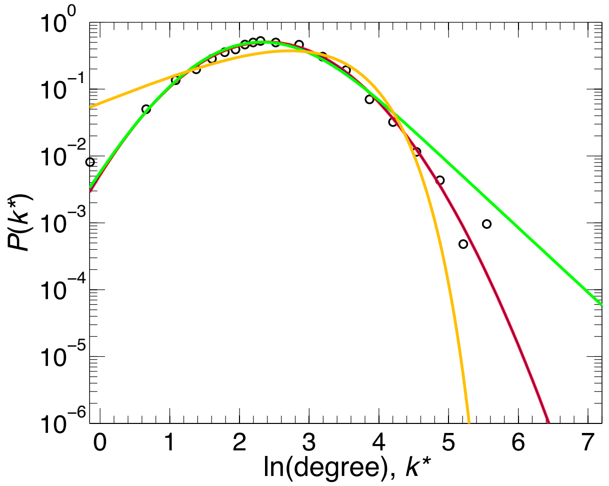

Figure S6. Degree distribution for the city with the highest coverage. Black circles: distribution of the regularly active callers (REC network, $\Delta T = 409$ days, $f_{\min}$ = 1/[90 days]). The Statistical City has $N = 17{,}535$ inhabitants and a total of $|S|$ = 12,304 assigned callers, resulting 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).

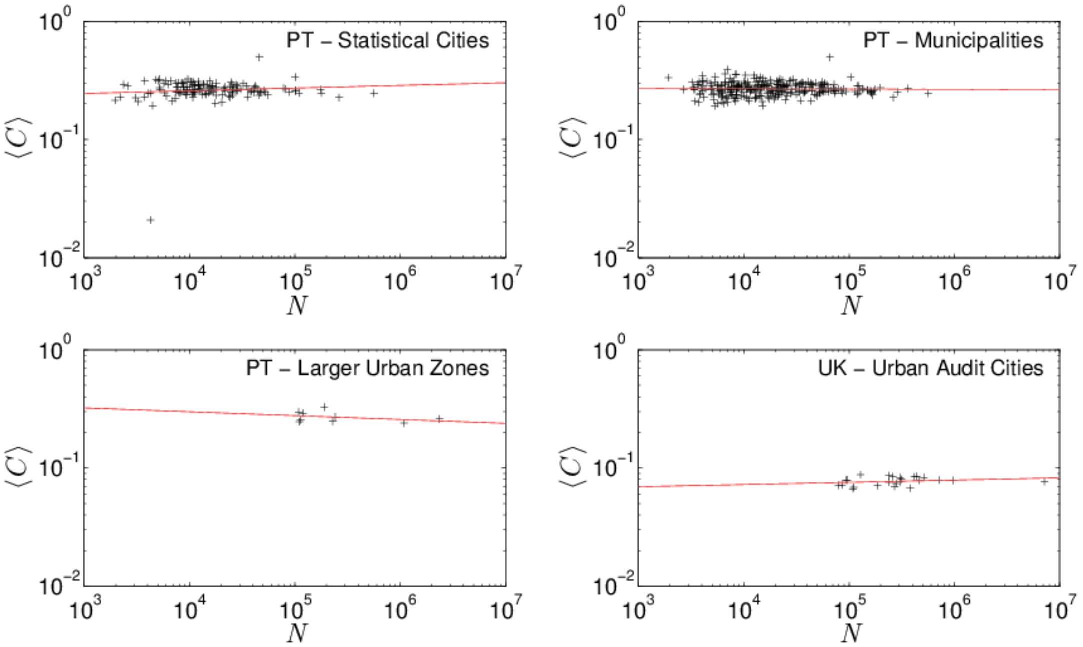

Figure S7. Regression analysis on the average clustering coefficients. Black crosses are the values of $\langle C \rangle$ versus city size for the different city definition as used in the main text. Red lines show the best linear regression, $\ln\langle C \rangle = \alpha + \gamma \ln N$. The values 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 $\text{Adj-}R^2 < 0.1$.

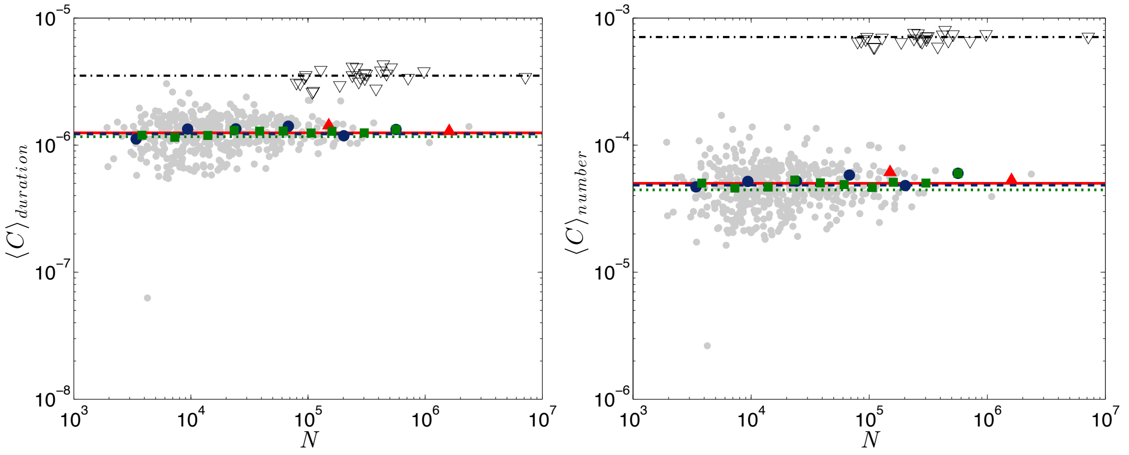

Figure S8. Average weighted clustering coefficients based on the call duration, $\langle C \rangle_{\text{duration}}$ (left), and based on the number of calls, $\langle C \rangle_{\text{number}}$ (right), for Portugal (REC network, $\Delta T = 409$ days) and UK, with symbols according to figure 3 in 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.

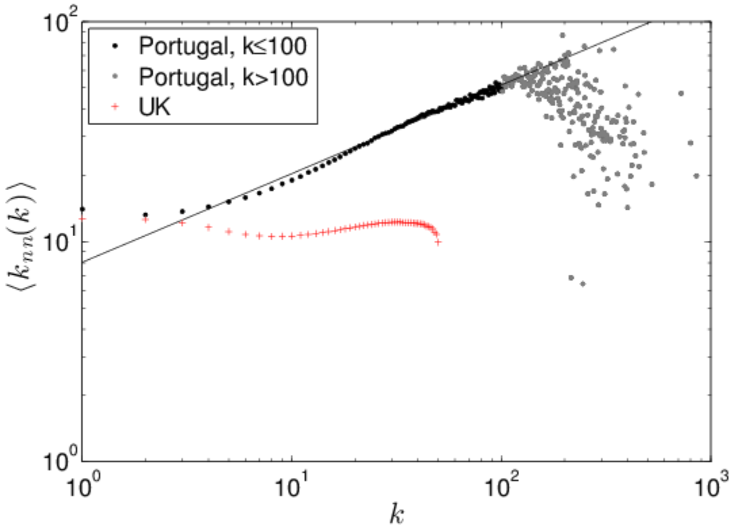

Figure S9. Average nearest-neighbour degree as a function of the nodal degree, $\langle k_{nn}(k) \rangle$. For Portugal (REC network, $\Delta T = 409$ days), the value varies like $\ln \langle k_{nn}(k) \rangle = \alpha + \gamma \ln k$ ($\gamma \approx 0.4$, Adj-$R^2 = 0.99$) for low values of k ($k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

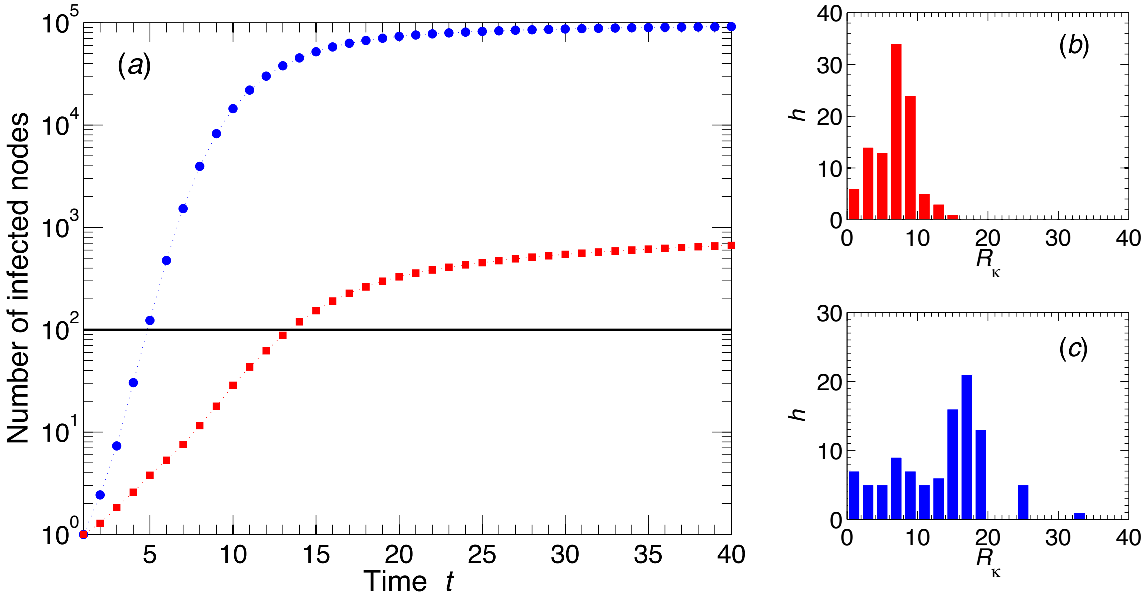

Figure S10. (*a*) The number of infected nodes as a function of the simulation time t for the examples of Lisbon (STC, with N=564,657 and $|S|$=109,448, blue circles), and Meda (STC, with N=2193, $|S|$=1033), averaged over 100 simulation runs (REC network, $\Delta T = 409$ days). Both urban units lie 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.

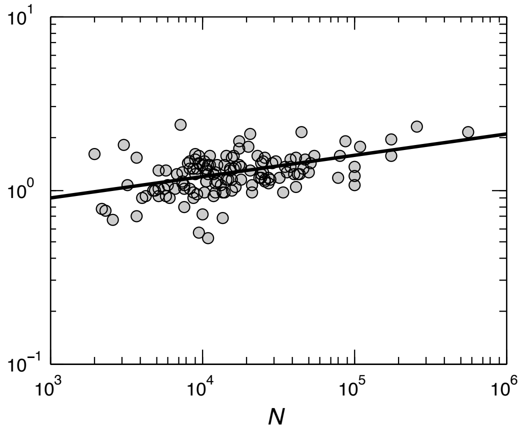

Figure S11. Spreading speed R for the Statistical Cities in Portugal based on the unweighted reciprocal network (REC network, $\Delta T = 409$ days). The spreading model has been implemented as for the weighted case (see main text), with the only difference that instead of the weight-dependent transmission 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 $s > s_{\text{min}}$	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, ΔT = 409 days, $f_{\min}$ = 1/[90 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.

			Distribution model			
City definition	No. of entities	Statistical method	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.

			Distribution model			
City definition	No. of entities	Statistical method	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.