The Origins of Scaling in Cities

Luís M. A. Bettencourt

Science **340**, 1438 (2013);
DOI: 10.1126/science.1235823

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

The Origins of Scaling in Cities

Luís M. A. Bettencourt

Despite the increasing importance of cities in human societies, our ability to understand them scientifically and manage them in practice has remained limited. The greatest difficulties to any scientific approach to cities have resulted from their many interdependent facets, as social, economic, infrastructural, and spatial complex systems that exist in similar but changing forms over a huge range of scales. Here, I show how all cities may evolve according to a small set of basic principles that operate locally. A theoretical framework was developed to predict the average social, spatial, and infrastructural properties of cities as a set of scaling relations that apply to all urban systems. Confirmation of these predictions was observed for thousands of cities worldwide, from many urban systems at different levels of development. Measures of urban efficiency, capturing the balance between socioeconomic outputs and infrastructural costs, were shown to be independent of city size and might be a useful means to evaluate urban planning strategies.

Cities exist, in recognizable but changing forms, over an enormous range of scales (*1*), from small towns with just a few people to the gigantic metropolis of Tokyo, with more than 35 million inhabitants. Many parallels have been suggested between cities and other complex systems, from river networks (*2*) and biological organisms (*3*–*6*) to insect colonies (*1*, *7*) and ecosystems (*8*). The central flaw of all these arguments is their emphasis on analogies of form rather than function, which limit their ability to help us understand and plan cities.

Recently, our increasing ability to collect and share data on many aspects of urban life has begun to supply us with better clues to the properties of cities, in terms of general statistical patterns of land use, urban infrastructure, and rates of socioeconomic activity (*6*, *9*–*13*). These empirical observations have been summarized across several disciplines, from geography to economics, in terms of how different urban quantities (such as the area of roads or wages paid) depend on city size, usually measured by its population, N.

The evidence from many empirical studies over the past 40 years points to there being no special size to cities, so that most urban properties, Y, vary continuously with population size and are well described mathematically on average by power-law scaling relations of the form $Y = Y_0 N^{\beta}$, where Y_0 and β are constants in N. The surprise, perhaps, is that cities of different sizes do have very different properties. Specifically, one generally observes that rates of social quantities (such as wages or new inventions) increase per capita with city size (*11*, *12*) (superlinear scaling, $\beta = 1 + \delta > 1$, with $\delta \simeq 0.15$), whereas the volume occupied by urban infrastructure per capita (roads, cables, etc.) decreases (sublinear scaling, $\beta = 1 - \delta < 1$) (Fig. 1). Thus, these data summarize familiar expectations that larger cities are not only more expensive and congested, but also more exciting and creative when compared to small towns.

These empirical results also suggest that, despite their apparent complexity, cities may actually be quite simple: Their average global properties may be set by just a few key parameters (*12*, *13*). However, the origin of these observed scaling relations and an explanation for the interdependences between spatial, infrastructural, and social facets of the city have remained a mystery.

Here, I develop a unified and quantitative framework to understand, at a theoretical level, how cities operate and how these interdependencies arise. Consider first the simplest model of a city with circumscribing land area A and

Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA.

E-mail: bettencourt@santafe.edu

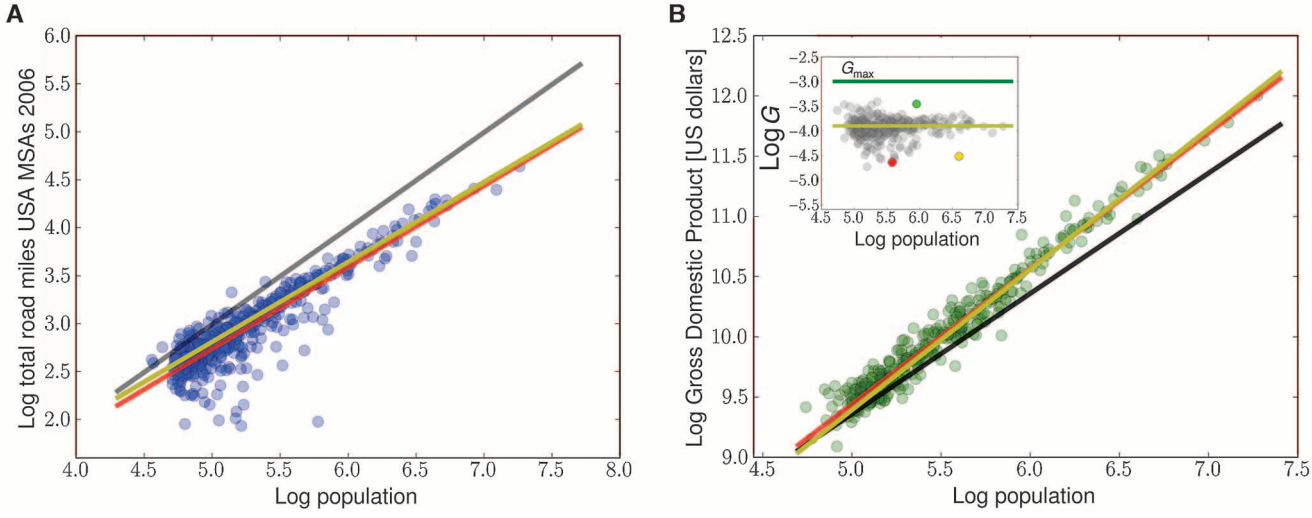


Fig. 1. Scaling of urban infrastructure and socioeconomic output. (A) Total lane miles (volume) of roads in U.S. metropolitan areas (MSAs) in 2006 (blue dots). Data for 415 urban areas were obtained from the Office of Highway Policy Information from the Federal Highway Administration (*14*). Lines show the best fit to a scaling relation $Y(N) = Y_0 N^{\beta}$ (red), with $\beta = 0.849 \pm 0.038$ [95% confidence interval (CI), $R^2 = 0.65$]; the theoretical prediction, $\beta = 5/6$ (yellow); and linear scaling $\beta = 1$ (black). **(B)** Gross metropolitan product of MSAs in 2006 (green dots). Data obtained for 363 MSAs from U.S. Bureau of Economic Analysis (*14*). Lines describe best fit (red) to data, $\beta = 1.126 \pm 0.023$ (95% CI, $R^2 = 0.96$); the theoretical prediction, $\beta = 7/6$ (yellow); and proportional scaling, $\beta = 1$ (black). The two best-fit parameters in each scaling relation were estimated by means of ordinary least-squares minimization to the linear relation between logarithmically transformed variables (*14*). The inset shows the estimate of G for 313 U.S. MSAs and the conservation law $\frac{d \ln G}{d \ln N} = 0$ ($R^2 = 0.003$). G is measured as the product of gross domestic product and road volume, both per capita. As predicted by the theory, observed values of G for different cities cluster around its most likely value (mode, yellow line), which gives an estimate of the optimum G^*, and are bounded by the maximum $G_{max} \simeq 8G^*$ (green line); see also Fig. 2B. Several metropolitan areas, such as Bridgeport, Connecticut (green circle); Riverside, California (yellow circle); or Brownsville, Texas (red circle), are outliers, suggesting that they are suboptimal in terms of their transportation efficiency or amount of social mixing.

population N. I write the interactions between people i, j in terms of a social network, F_k^{ij}, and assume that social interactions [e.g., friendship, employment, acquaintance, etc.] are local, take place over an interaction area a_0 (a cross section in the language of physics), and have strength g_k, where k describes social link types (*14*). The parameters, g_k, can be either positive (attractive, expressing a social benefit, e.g., mutually beneficial economic relations) or negative (repulsive, expressing a social cost, e.g., crime). All these processes share the same average underlying dynamics of social encounters in space and time, against the background of the city and its infrastructure networks.

The average number of local interactions per person is given by the product of the volume spanned by their movement, $a_0\ell$, times the population density $n = N/A$, where ℓ is the typical length traveled by people, goods, and information (*14*). The total average social output of a city can be obtained by multiplying the total number of interactions by the average outcome per interaction, $\overline{g}$, leading to $Y = G\frac{N^2}{A}$, with the parameter $G \equiv \overline{g}a_0\ell$ measuring the product of average social output times area, both per capita (Fig. 1). Each urban socioeconomic output, Y, has physical units set by g_k, but it is useful to think of all quantities ultimately expressed in terms of energy per unit time (power).

Another crucial property of cities is that they are mixing populations. That is, even if people in the city explore different locations at different times, anyone can in principle be reached by anyone else. This concept, familiar from population biology (*15*), is the basis of definitions of functional cities as metropolitan statistical areas (MSAs), e.g., by the U.S. census bureau. In practice, this means that the cost per person of a mixing population is proportional to the transverse dimension (diameter), L, of the city $L \sim A^{1/2}$. Thus, the total power spent in transport processes to keep the city mixed is $T = \varepsilon LN = \varepsilon A^{1/2}N$, where ε is a force per unit time. This cost must be covered by each individual's budget, $y = Y/N$, requiring $y \simeq T/N$, which implies $A(N) = aN^{\alpha}$ with $\alpha = 2/3$ and $a = (G/\varepsilon)^{\alpha}$. The baseline area, a, increases with more productive interactions, e.g., due to economic growth, and decreasing transportation costs, as is observed in worldwide patterns of urban sprawl over time (*16*). Thus, I obtain $Y = Y_0N^{\beta}$, where $\beta = 2 - \alpha = 1 + 1/3 > 1$ and $Y_0 = G^{1-\alpha}\varepsilon^{\alpha}$. This simple model leads to area, A, varying sublinearly with N ($\alpha = 2/3 < 1$), and socioeconomic outputs, Y, varying superlinearly ($\beta = 4/3 > 1$). However, this overestimates β because as cities grow, space becomes occupied and transportation of people, goods, and information is channeled into networks. The space created by these networks gives the correct measure of the social interactions that can occur in cities.

I propose a more realistic model by generalizing these ideas in terms of four simple assumptions:

1) Mixing population. The city develops so that citizens can explore it fully given the resources at their disposal. I formalize this principle as an entry condition (*17*), by requiring that the minimum resources accessible to each urbanite, $Y_{\min}/N \sim GN/A$, match the cost of reaching anywhere in the city. Because travel paths need not be linear, I generalize their geometry via a fractal dimension, H, so that distance travelled $\propto A^{H/D}$ (*14*). Matching interaction density to costs, I obtain a generalized area scaling relation, $A(N) = aN^{\alpha}$, with a as before and $\alpha = \frac{2}{2+H}$ [$\alpha = \frac{D}{D+H}$ in D dimensions]. $H = 1$ allows individuals to fully explore the city within the smallest distance traveled, implying that N scales like a physical volume (*14*, *18*).

2) Incremental network growth. This assumption requires that infrastructure networks develop gradually to connect people as they join, leading to decentralized networks (*6*, *19*). Specifically, the scaling of Fig. 1A is obtained when the average distance between individuals $d = n^{-1/2} = (A/N)^{1/2}$ equals the average length of infrastructure network per capita so that the total network area, $A_n(N) \sim Nd = A^{1/2}N^{1/2}$. Together with the first assumption, this implies that $A_n \sim a^{1/2}N^{1-\delta}$ with $\delta = 1/6$ [$A_n \sim A^{1/D}N^{(D-1)/D} = a^{1/D}N^{1-\delta}$, with $\delta = \frac{H}{D(D+H)}$ in D dimensions]. This has been observed in U.S. and German road networks (*6*, *12*, *19*) and tracks the average built area of more than 3600 large cities worldwide (*16*), measured through remote sensing.

3) Human effort is bounded, which requires that G is, on average, independent of N, i.e., $dG/dN = 0$ (Fig. 1B, inset). The increasing mental and physical effort that growing cities can demand from their inhabitants has been a pervasive concern to social scientists (*20*). Thus, this assumption is necessary to lift an important objection to any conceptualization of cities as scale-invariant systems. Bounded effort is also observed in urban cell phone communication networks (*21*) and is in general a function of human constraints and urban services and structure.

4) Socioeconomic outputs are proportional to local social interactions, so that $Y = GN^2/A_n \sim N^{1+\delta}$. From this perspective, cities are concentrations not just of people, but rather of social interactions. This point was emphasized by Jacobs (*22*, *23*), but has been difficult to quantify. The prediction that social interactions scale with $\beta = 1 + \delta \simeq 7/6$ was observed recently in urban telecommunication networks (*21*). Together these assumptions predict scaling exponents for a wide variety of urban indicators, from patterns of human behavior and properties of infrastructure to the price of land (*6*, *9*–*12*, *16*, *21*, *24*, *25*), summarized in Table 1 (*14*).

Thus far, I obtained estimates for scaling exponents without the need for a detailed model of infrastructure. Next, I show how network models of infrastructure can help to illuminate urban planning issues. Consider the infrastructure in a city described by a network with h hierarchical levels (Fig. 2A). The network branching, b, measures the average ratio of the number of units of infrastructure at successive levels, $N_i = b^i$, e.g., number of paths to small roads, or larger roads to highways. I assume that the number of infrastructure units at the lowest level, $i = h$, equals the number of people, so that $N_h = N$ and $h = \ln N/\ln b$. These networks are not hierarchical trees (*26*) (Fig. 2A). The length of a network segment (such as a road) at level i is l_i, crossing a land area a_i, and its transverse dimension is s_i, an area in 3D networks and a length in 2D. To obtain the above scaling relations, I assume that the transverse dimension of the smallest network units, s_*, is independent of N. This leads to the scaling of network width, $s_i = s_* b^{(1-\delta)(h-i)}$, which says that highways or water mains are much wider than building corridors or household

Table 1. Urban indicators and their scaling relations. Columns show measured exponent ranges (see table S3 for details). Also shown are predicted values for $D = 2$, $H = 1$ (the simplest theoretical expectation) and for general D, H. Agglomeration effects vanish as $H \to 0$ (*14*). The larger range for the observed land-area exponent is likely the result of different definitions of the city in space and distinct measurement types. See table S3 and supplementary text for specific values of observed exponents, discussion, and additional data sources.

Urban scaling relations	Observed exponent range	Model ($D = 2$, $H = 1$)	Model D, H
Land area $A = aN^{\alpha}$	[0.56,1.04]	$\alpha = \frac{2}{3}$	$\alpha = \frac{D}{D+H}$
Network volume $A_n = A_0N^{\nu}$	[0.74,0.92]	$\nu = \frac{5}{6}$	$\nu = 1 - \delta$
Network length $L_n = L_0N^{\lambda}$	[0.55,0.78]	$\lambda = \frac{2}{3}$	$\lambda = \alpha$
Interactions per capita $\bar{I}_i = I_0N^{\delta}$	[0.00,0.25]	$\delta = \frac{1}{6}$	$\delta = \frac{H}{D(D+H)}$
Socioeconomic rates $Y = Y_0N^{\beta}$	[1.01,1.33]	$\beta = \frac{7}{6}$	$\beta = 1 + \delta$
Network power dissipation $W = W_0N^{\omega}$	[1.05,1.17]	$\omega = \frac{7}{6}$	$\omega = 1 + \delta$
Average land rents $P_L = P_0N^{\delta_L}$	[0.46,0.52]	$\delta_L = \frac{1}{2}$	$\delta_L = 1 - \alpha + \delta$

pipes, $s_0 = s_* b^{(1-\delta)h} \gg s_h = s_*$. Additionally, because infrastructure must reach everyone in the city (*6*, *18*), total network length is area filling, $l_i = a_i/l$,with $a_i = ab^{(\alpha-1)i}$. This means that the land area per person, $a_h = aN^{\alpha-1}$, and shortest network distance, $l_h = (a/l)N^{\alpha-1}$, which defines l, decrease with N. The total network length L_n and network area A_n follow from the sum of the geometric series over levels

$$L_n = \sum_{i=0}^{h} l_i N_i = \frac{a}{l}\sum_{i=0}^{h} b^{\alpha i} = \frac{a}{l}\frac{b^{\alpha(h+1)}-1}{b^{\alpha}-1} \simeq L_0 N^{\alpha}, L_0 = a/l \quad (1)$$

$$A_n = \sum_{i=0}^{h} s_i l_i N_i = s_* \frac{a}{l} b^{(1-\delta)h} \sum_{i=0}^{h} b^{(\alpha+\delta-1)i} \simeq A_0 N^{1-\delta}, \quad A_0 = \frac{s_* a}{l(1-b^{\alpha+\delta-1})} \quad (2)$$

where I took $\alpha + \delta < 1$, which holds for $D > 1$.

I can now compute the cost of maintaining the city connected as the energy necessary for moving people, goods, and information across its infrastructure networks. These movements form a set of currents, transporting various quantities across the city and can be quantified by means of the language of circuits. The scaling of s_i together with total current, J, conservation across levels $J_i = s_i \rho_i v_i N_i = s_{i-1}\rho_{i-1} v_{i-1} N_{i-1} = J_{i-1}$ for all i, sets the scaling for $\rho_i v_i$, the current density at level i, where ρ_i is the density of carriers in the network and v_i their average velocity. This quantity is interesting because it controls the dissipation mechanisms in any network. I obtain $\rho_i v_i = b^{-\delta}\rho_{i-1} v_{i-1}$, which implies that the current density decreases with increasing i, so that highways are faster and/or more densely packed than smaller roads (*27*, *28*). Making the additional assumption that individual needs, $\rho_h v_h = \rho_* v_*$, are independent of N (*12*) leads to $\rho_i v_i = b^{\delta(h-i)}\rho_* v_*$. Then, the total current $J_i = J = J_0 N$, with $J_0 = s_* \rho_* v_*$, which is a function only of individuals' characteristics.

There are many forms of energy dissipation in networks, including those that occur at large velocity or density. Here, I make the standard assumption that the resistance per unit length per transverse network area, r, is constant (*2*, *5*), leading to the resistance per network segment, $r_i = r\frac{l_i}{s_i}$. For N_i parallel resistors this gives the total resistance per level, $R_i = \frac{r_i}{N_i} = \frac{ar}{ls_*} b^{-(1-\alpha+\delta)i-(1-\delta)h}$. The total power dissipated, W, follows from summing $W_i = R_i J_i^2$ over levels,

$$W = J^2 \sum_{i=1}^{h} R_i = J^2 \frac{ar}{ls_*} b^{-(1-\delta)h} \frac{1-b^{-(1-\alpha+\delta)(h+1)}}{1-b^{-1+\alpha-\delta}} \simeq W_0 N^{1+\delta}, \quad W_0 = \frac{arJ_0^2}{ls_*(1-b^{-1+\alpha-\delta})} \quad (3)$$

which scales superlinearly, with exponent $1+\delta = 1+1/6$ in $D = 2$, $H = 1$. Thus, energy dissipation scales with population like social interactions, as observed in German urban power grids (*12*), so that the ratio Y/W, a measure of urban efficiency, is independent of city size.

Finally, I show that these results can be derived by maximizing net urban output, $\mathcal{L}$, as the difference between social interaction outcomes, Y, and infrastructure energy dissipation, W, under settlement and network constraints,

$$\mathcal{L} = Y - W + \lambda_1(\varepsilon A^{H/D} - GN/A) + \lambda_2(A_n - cNd) \underset{d\mathcal{L}/dG=0}{\rightarrow} \frac{2\alpha-1}{\alpha} G^* \frac{N^2}{A_n(N)} \quad (4)$$

where $c = A_0 a^{-1/D}$ and λ_1, λ_2 are Lagrange multipliers. Equation 4 gives the basis for the derivation of the properties of every segment in the network, through Eqs. 1 and 2, in analogy with (*2*, *4*, *5*). The novelty in Eq. 4 is the prediction of an optimal $G = G^*$, through $d\mathcal{L}/dG = 0$, and the expectation that values of G for different cities fluctuate around this value, as observed in Fig. 1B (inset).

To see this, consider that, keeping ε fixed and $a = (G/\varepsilon)^{\alpha}$, both $Y$and W grow with G, because $Y_0 \sim G^{1-\alpha}$ and $W_0 \sim G^{\alpha}$. This tension between social interactivity, transportation costs, and spatial settlement patterns is at the root of most urban planning and policy. The limiting values of G follow from the solutions to $\mathcal{L} = 0 : G = 0$ and $G = G_{\max} = \left[\frac{(\varepsilon l)^{2\alpha}}{r' J_0^2} l^{2(1-\alpha)}\right]^{\frac{1}{2\alpha-1}}$, where $r' \approx r$ (*14*). It follows that $G^* = \left(\frac{1-\alpha}{\alpha}\right)^{1/(2\alpha-1)} G_{\max} \simeq G_{\max}/8$, with $\alpha \simeq 2/3$ (Fig. 1B, inset). Thus, cities will form if the balance of social interactions is positive, $\bar{g} > 0$. However, there is an upper value of $G = G_{\max}$(Fig. 1B, inset) beyond which dissipation costs overcome social benefits and a city may split up into regions. For $G < G^*$, the social interaction potential of a city is underdeveloped. Such places tend to be poorer and have less advanced infrastructure. Thus, I would expect that cities such as Riverside, California, or Brownsville, Texas (Fig. 1B), where estimates of G are less than average, would typically benefit from measures

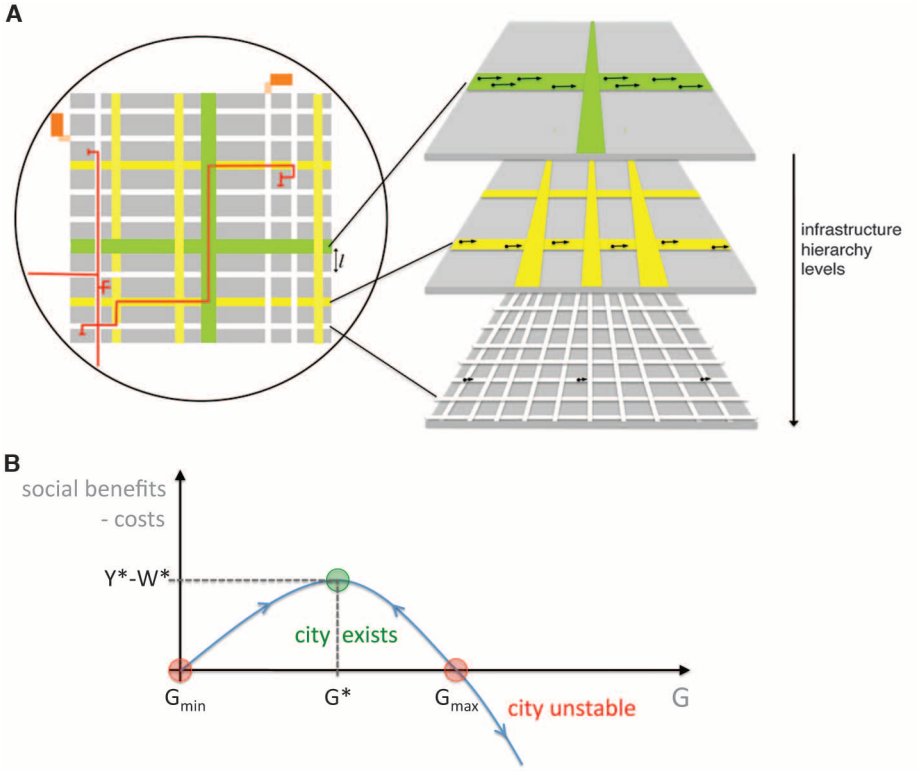


Fig. 2. The spatial city and its social and dissipative processes. (**A**) Gray blocks denote settled areas, and spaces in between (white, yellow, green) represent infrastructure networks, treated in terms of a size hierarchy. Total network length $L_n = 2(n_b + 1)L \simeq A/l$ is area filling (circle), where n_b is the number of blocks across the city (*14*). Red lines denote the volume of public space spanned by an individual, which determines his or her average number of social interactions. As the city grows and new land is settled (orange blocks), the infrastructure network grows incrementally (orange segments). The flux $\rho_i v_i$ in larger network segments is higher (black dots plus arrows), controlling the energy dissipation in the city. (**B**) There is an optimal value of G at which cities are most productive. Cities can exist when social interactions are positive $G > G_{\min} = 0$, and less than an upper value $G < G_{\max}$ (red circles), at which point dissipation costs overcome benefits. The optimal $G = G^*$ (green circle) corresponds to the most efficient city.

that promote greater mobility or density, in order to achieve more intense and beneficial city-wide social contact. Conversely, cities with $G > G^*$ become victims of their socioeconomic success by incurring escalating mobility costs. Bridgeport, Connecticut's, MSA (Fig. 1B) may be developed in terms of its economic functions and infrastructure, but might generally benefit from more compact urban living or from increases in transportation energy efficiency. That is, cities may be suboptimal either because they do not realize their full social potential or because they do so in a manner that renders transportation costs too high. In either case, this approach shows how urban planning must take into account the delicate net balance between density, mobility, and social connectivity and thus provides a general framework for the iterative development and assessment of urban policies.

That many cities are becoming more global in their economic relations and political and cultural influence (*29*) does not alter the basic premises of the theory. The internal dynamics and organization of cities (as social networks of people and institutions) produces new socioeconomic functions that allow cities to exchange goods, services, people, and information within and across national borders (*22*, *23*, *30*). Thus, even if some singular places such as Hong Kong, Singapore, or Dubai are primarily part of international economies, the majority of the world's most global cities, such as Tokyo, New York, Los Angeles, Beijing, Shanghai, Berlin, or Frankfurt, show clear scaling effects in line with their own national urban systems (Fig. 1 and figs. S1 to S3).

All cities have spatial and social pockets of greater and lower mobility, social integration, better or worse services, and so forth (*1*, *17*). It should be emphasized that the theory does not predict density profiles or socioeconomic differences inside the city, but the scaling for the properties of the city as a whole. None of these pockets exist in absolute isolation; they are just more or less "connected," so they must be understood with reference to the rest of the city (*17*).

The interactions between people also provide the basis for institutional relationships via the appropriate groupings of individuals in social or economic organizations and by the consideration of the resulting links between such entities. Institutions and industries that benefit from strong mutual interactions may aggregate in space and time within the city in order to maximize their $Y - W$, a point first made by Marshall (*23*) in the context of industrial districts. Other organizations may benefit primarily from the general effects that result from being in the wider city and collecting a diversity of interactions, an argument often attributed to Jacobs (*22*). These results establish necessary conditions for urban areas to express certain levels of socioeconomic productivity, but it remains a statistical question (*21*, *25*) how well they are realized in specific places.

Most urban systems for which reliable data exist confirm almost exactly the simplest predictions of the theory developed here. Examples are the scaling of area for about 1800 cities in Sweden (*14*, *18*), or for roads in several hundred American (Fig. 1A) and Japanese metropolitan areas (fig. S3). One of the most spectacular agreements is for the scaling of total area of paved surfaces for all cities worldwide above 100,000 people (over 3600 cities) (*14*, *16*). These examples illustrate the result derived above that urban infrastructure volume scales faster with population than land area (and both are sublinear). This effect is visually apparent in large, developed cities, where roads, cables, and pipes become ubiquitous and eventually migrate into the third dimension, above or below ground.

Measurements of electrical cable length and dissipative losses in German urban power grids (*12*) further confirm these expectations and support another key result obtained above: The energy loss in transport processes scales like socioeconomic rates (and both are superlinear). This shows how cities are fundamentally different from other complex systems, such as biological organisms (*4*, *5*) or river networks (*2*), which are thought to have evolved to minimize energy dissipation. Thus, the framework developed here also brings into focus efforts for sustainable urban development, by showing what kind of energy budget must be expended in order to keep cities of varying sizes socially connected.

The predictions of the theory are further supported by data on the size of urban economies from hundreds of cities in several continents, such as those in the United States (Fig. 1B), Japan (fig. S3), China (fig. S2A), or Germany (fig. S2B). In particular, the specific result that scaling exponents remain invariant over time, and are independent of population size and level of development, is confirmed by data for wages in U.S. metropolitan areas spanning 40 years (fig. S3). Direct empirical tests on the predictions made here for individual properties remain more difficult, but are confirmed, for example, by measurements for the scaling of social interactions with city size in the cell phone networks of two European nations (*21*), and for certain other patterns of individual behavior (*12*, *20*, *31*). Nevertheless, for most nations, we cannot yet access all predicted urban quantities simultaneously, especially in developing countries. This provides many future tests and applications for the theory, especially where understanding urbanization is most critical.

The spatial concentration and temporal acceleration of social interactions in cities has some striking qualitative parallels in other systems that are also driven by attractive forces and become denser with scale (*20*, *30*). The most familiar are stars, which burn faster and brighter (superlinearly) with increasing mass. Thus, although the form of cities may resemble the vasculature of river networks or biological organisms, their primary function is as open-ended social reactors. This view of cities as multiple interconnected networks that become denser with increasing scale (*32*) may also help to elucidate the function of other systems with similar properties, from ecosystems to technological information networks, despite their different relationships to physical space.

References and Notes

1. L. Mumford, *The City in History: Its Origins, Its Transformations, and Its Prospects* (Harcourt, Brace & World, Inc., New York, 1961).
2. I. Rodríguez-Iturbe, A. Rinaldo, *Fractal River Basins: Chance and Self-Organization* (Cambridge Univ. Press, New York, 1997).
3. E. N. Bacon, *The Design of Cities* (Penguin Group USA, New York, 1976).
4. G. B. West, J. H. Brown, B. J. Enquist, *Science* **276**, 122 (1997).
5. G. B. West, J. H. Brown, B. J. Enquist, *Science* **284**, 1677 (1999).
6. H. Samaniego, M. E. Moses, *J. Transport. Land Use* **1**, 21 (summer 2008).
7. J. S. Waters, C. T. Holbrook, J. H. Fewell, J. F. Harrison, *Am. Nat.* **176**, 501 (2010).
8. D. S. Dendrinos, H. Mullally, *Urban Evolution: Studies in the Mathematical Ecology of Cities* (Oxford Univ. Press, Oxford, 1985).
9. L. Sveikauskas, *Q. J. Econ.* **89**, 393 (1975).
10. E. L. Glaeser, B. Sacerdote, *J. Polit. Econ.* **107**, S225 (1999).
11. L. M. A. Bettencourt, J. Lobo, D. Strumsky, *Res. Policy* **36**, 107 (2007).
12. L. M. A. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, G. B. West, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7301 (2007).
13. M. Batty, *Science* **319**, 769 (2008).
14. See supplementary materials for details.
15. R. Anderson, R. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Oxford, 1991).
16. S. Angel, J. Parent, D. L. Civco, A. Blei, D. Potere, *Prog. Plann.* **75**, 53 (2011).
17. D. Saunders, *Arrival City* (Pantheon Books, New York, 2010).
18. S. Nordbeck, *Geogr. Ann.* **53**, 54 (1971).
19. M. Barthélemy, *Phys. Rep.* **499**, 1 (2011).
20. S. Milgram, *Science* **167**, 1461 (1970).
21. M. Schläpfer *et al.*, *The Scaling of Human Interactions with City Size*; http://arxiv.org/abs/1210.5215.
22. J. Jacobs, *The Economy of Cities* (Vintage Books, New York, 1969).
23. A. Marshall, *Principles of Economics* (MacMillan, London, 1920). See book IV, chapter X.
24. E. L. Glaeser, J. D. Gottlieb, *J. Econ. Lit.* **47**, 983 (2009).
25. A. Gomez-Lievano, H. Youn, L. M. A. Bettencourt, *PLoS ONE* **7**, e40393 (2012).
26. C. Alexander, *Archit. Forum* **122**, 58 (1965).
27. A. Downs, *Traffic Quarterly* **16**, 393 409 (1962).
28. G. Duranton, M. A. Turner, *Am. Econ. Rev.* **101**, 2616 (2011).
29. S. Sassen, *The Global City: New York, London, Tokyo* (Princeton Univ. Press, Princeton NJ, 1991).
30. M. Fujita, P. Krugman, A. J. Venables, *The Spatial Economy: Cities, Regions, and International Trade* (MIT Press, Cambridge MA, 2001).
31. F. Lederbogen *et al.*, *Nature* **474**, 498 (2011).
32. J. Leskovec, J. Kleinberg, C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations." In *KDD '05 Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 177–187 (2005); http://dl.acm.org/citation.cfm?id=1081893.

Acknowledgments: I thank J. Lobo, G. West, and H. Youn for discussions. This research was supported by grants from the Rockefeller Foundation, James S. McDonnell Foundation (grant 220020195), NSF (grant 103522), Bill and Melinda Gates Foundation (grant OPP1076282), John Templeton Foundation (grant 15705), and Bryan J. and June B. Zwan Foundation.

Supplementary Materials

www.sciencemag.org/cgi/content/full/340/6139/1438/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S3
Tables S1 to S3
References (*33*–*63*)

29 January 2013; accepted 29 April 2013
10.1126/science.1235823

www.sciencemag.org/cgi/content/full/340/6139/1438/DC1

Supplementary Materials for

The Origins of Scaling in Cities

Luís M. A. Bettencourt

E-mail: bettencourt@santafe.edu

Published 21 June 2013, *Science* **340**, 1438 (2013)
DOI: 10.1126/science.1235823

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S3
Tables S1 to S3
References (*33*–*63*)

Materials and Metods

Data Sources

Data for road surface in United States Federal-Aid Urbanized Areas (mostly equivalent to Metropolitan Statistical Areas, or MSAs), shown in Fig. 1A, is provided by the Office of Highway Policy Information from the Federal Highway Administration. It is available online at `http://www.fhwa.dot.gov/policyinformation/statistics.cfm`, under "Urbanized Area Summaries". Data from the 2006 report was used in Fig. 1A. Gross Metropolitan Product for US Metropolitan Statistical Areas, shown in Fig. 1B, is compiled by the US Bureau of Economic Analysis (BEA) since 2001 and is available online at `http://www.bea.gov/regional/`. Data for 2006 was used to produce Fig. 1B.

Data for the supplementary figures was obtained from Urban Audit/Eurostat (available at `http://www.urbanaudit.org`), from collections of official Chinese statistics (available at: `http://chinadataonline.org`), from Japan's Statistical Yearbook (available at `http://www.stat.go.jp/english/data/`) and from the US Bureau of Economic Analysis (BEA) (available at `http://www.bea.gov/regional/`). The year or temporal period used in producing Figs. S1-3 is indicated in the corresponding caption.

Estimation of Scaling Parameters

Best fits to data to obtain the scaling parameters in Figs. 1A, 1B, S1-3, were performed using ordinary least squares minimization to a linear relation of the logarithmically transformed variables. These fits involve the estimation of two parameters: a pre-factor (such as Y_0), and an exponent, such as β (see Figs. 1, S1-3). The number of cities (taken as Metropolitan Areas or their equivalent) for which data are available varies with the size of each national urban system, between several hundred (in the US and China) and tens of cities. e.g. for Japan or Germany, see Table S3. For some quantities, data are available for thousands of cities, either by aggregating

across national urban systems (*16*) or by taking into account smaller population agglomerations (*18*,*25*).

Parameter Definitions and Estimated Values

Table S1 gives a summary of fundamental urban input variables and of key, theoretically derived, urban parameters. Table S2 shows the explicit dependence of the scaling pre-factors and exponents on these parameters, as derived in the main text and presented in Table 1. Table S3 provides an annotated synthesis of data sources and of the existing literature on empirical urban scaling relations (including new results reported here). Table S3 also provides a general overview of estimated values for exponents and their uncertainty intervals, summarized in Table 1. Additional discussion of data sources, the nature of proxy quantities and specific issues relating to each scaling relation is provided in Supplementary Text, below.

Supplementary Text

From Social Interaction Networks to Average Socioeconomic Rates

In general, I describe social interactions in a city in terms of a generalized graph, F_{ij}^k, (a graph between elements i and j, mediated by a set of different interaction types - friendship, employment, acquaintance, etc - indexed by k) as

$$Y = \sum_{i,j;k} g_k F_{ij}^k, \tag{S1}$$

where g_k is the strength per link of the interaction of type k to generate the total output of the city, Y. Note that the couplings, g_k, can be either positive (attractive, expressing a social benefit, e.g. mutually beneficial economic relations) or negative (repulsive, expressing a social cost e.g. crime), though the balance must be positive for the city to exist, see below. The couplings g_k have dimensions of Y per interaction, for example units of money or energy per

unit time, per interaction. In a city there are many forms of interactions. For example, economic transactions contribute to economic output in terms of wages, profits, and many other quantities. Crime, in contrast, may be the output of non-economic interactions such as those between the perpetrator and the victim as well as those mediated by law enforcement and by citizens themselves. Likewise, the interactions that lead to the spread of a contagious disease will be mediated by their specific types of encounters. The urban environment affects its citizens across all these dimensions so that a theory of cities must take them into account *together*.

The essential point I make here is that all these processes share the same average underlying dynamics of social encounters in space and time, against the background of the city and its infrastructure networks. To see this more explicitly, first consider the number of interactions, $I_{i,k}$, of a specific individual i, across all modes, k,

$$I_{i,k} = \sum_j F_{ij}^k. \tag{S2}$$

I consider the situation where the strength of the interaction k, is statistically independent of the specific pair i, j so that I can write $F_{ij}^k = p(k|ij)F_{ij} = p(k)F_{ij}$, where $p(k)$ is the probability of different interaction modes, k, per link and F_{ij} is the social network across all interaction types. Now consider that these interactions take place in space and time. Each individual is characterized by an interaction area, a_0, (a cross section in the language of physics) and by a length traveled in the network, ℓ. This spans a *worldsheet*, which is a fraction of the total public space volume, V_n, (or area, in 2D networks, A_n, which is the notation used in the main text) of the city. Because both a_0 and ℓ are intrinsic properties of individuals I take these two parameters as independent of the type of interactions k.

Taking all people to be homogeneously distributed in this volume (the mean field assumption), the *average* total interactions experienced by our test individual, $\bar{I}_{i,k}$, are given by the

ratio of the two volumes times the total number of (other) individuals, i.e.

$$\bar{I}_{i,k} = p(k) \int d^D x \Gamma(x) \delta\left(x - x(t)\right) \simeq p(k) \Gamma_n \int dx^D \delta\left(x - x(t)\right)$$
$$= p(k) a_0 \ell \frac{N-1}{V_n} \simeq p(k) a_0 \ell \frac{N}{V_n}, \tag{S3}$$

with $\Gamma(x)$ the density in public networks (and Γ_n its average value), where interactions take place. In this last expression I wrote $N - 1 \simeq N$, for large N (*33*). Note that any other sufficiently short-range interaction potential, not necessarily a δ-function, would lead to the same result, Eq. (S3), up to a dimensionless multiplicative constant, independent of N. Then, I can finally write

$$Y = \sum_k \bar{I}_{i,k} = \bar{g} \frac{a_0 \ell}{V_n} N, \qquad \bar{g} = \sum_k p(k) g_k, \tag{S4}$$

which are the total interactions experienced on the average by an individual i, in a city of population N, and public volume V_n. The coupling $\bar{g}$ is the average strength per link of interactions over all modes. In the main text, I take the volume of public space of the city to scale like that of its infrastructure networks, A_n. In the two dimensional case (D=2) the cross sectional area a_0 takes the dimension of a traverse length, so that the ratio of (2-dimensional) volumes remains a pure dimensionless number. Thus, we obtain

$$Y = G \frac{N^2}{A_n}, \tag{S5}$$

with $G \equiv \bar{g} a_0 \ell$.

It is important to stress that although social interactions are local and take place at the most microscopic level between two individuals, Eq. (S1) leads nevertheless to *effective* interactions between individuals that are not directly connected, through chains of people between them, and between individuals and institutions (firms, public administration) as well as between institutions themselves. These effective interactions are obtained via the appropriate groupings of individuals in social or economic organizations and by the consideration of the resulting coarse-grained interactions between such entities (which are always ultimately mediated by people).

Institutions and industries that benefit from strong mutual interactions may aggregate in space and time within the city in order to maximize their $Y - W$ (see below), a point first made by Marshall (*23*) in the context of industrial districts. Others organizations may benefit primarily from the mean-field effects that result from being in the wider city and collecting a diversity of interactions, an argument often attributed to Jane Jacobs (*22*). This analysis of the finer structure of more heterogeneous interactions, which requires considerations beyond the average behavior derived here, will be considered elsewhere. Likewise the analysis of the fine structure of types of functions and interactions in cities, for example in terms of professions, and their connection to superlinear increases in socioeconomic productivity is developed in greater detail in (*34*).

Mixing, Exploration of Space and Fractal Dimension

Here, I develop more detailed considerations about the exploration of space by individuals that may take place in cities and the necessary conditions for a mixing population. The general idea is that, to benefit from their integration in the city, individuals explore different locations at different times but must be able, on their most basic budget, to explore the city fully. I parameterize this general behavior in Eq. (S6) by H, the fractal (Hausdorff) dimension of a path in space. T is the (energetic) cost associated with such path, which is written in terms of the city's land area, A, as

$$T = \epsilon A^{\frac{H}{D}}, \tag{S6}$$

in D general dimensions and where ϵ is a cost per unit length (see main text). The minimum budget that a new citizen may naturally muster is $y_{\min} = GN/A$, which is much smaller than the average budget, $y = GN/A_n$, because $A >> A_n$. Thus, this can be seen as an entry condition into the city: A new citizen, perceiving the city only in an unstructured way, before knowing its networks and public spaces, should be able to reach anyone else in the city. Equating T to $y_{\min}$

leads to a generalized relationship between population and area of the form

$$A = aN^{\alpha}, \tag{S7}$$

with the exponent $\alpha = D/(D+H) \simeq 2/3$, for $D = 2$, $H = 1$, and the baseline area $a = (G/\epsilon)^{\alpha}$. Note that a is a rising function of G, which controls the average strength (and productivity) of social interactions, and of decreasing ϵ the cost of transport per unit length. Thus, increases in human capital, mobility and the diversity of social interactions, if expressed in increasing values of G and increases in transportation efficiency (decreases in ϵ), lead to a larger a and an overall less dense city, while preserving the scaling relation. This is consistent with the observed trend in modern cities over time to become less dense (*35*).

The parameters G and ϵ are generally time dependent and may also show some (small) city size dependences a subject that I explore in the main text. Note that G can be measured from $G = ya_n = \frac{Y}{N}\frac{A_n}{N}$, as is done in Figure 1B (inset). The constancy of the average G, measured in this way, gives direct evidence for the empirical validity of the assumptions made in the main text and can be interpreted as a "conservation law", since $dG/dN = 0$, or equivalently $d\ln G/d\ln N = 0$.

I introduced a Hausdorff dimension H to characterize paths through the city because I considered it too strong an assumption to take their geometry to be known. It is interesting to discuss the meaning of the several values of H further. $H = 1$ corresponds to the most natural assumption, that these costs are proportional to the linear extent (diameter) of the city and is clearly sufficient for an individual to reach any location in the city by himself. This is the assumption that is made (often implicitly) in urban economic models of land use, due to Alonso, Müth, Mills and others (*36, 37*). The exponent $\alpha = 2/3$ was derived long ago by Nordbeck (*18*), who first observed it for Swedish cities. He used an allometric argument that total city population, N, should scale like a 3D volume due to its spatial profile of density change, which implies

$N \sim A^{3/2}$. This is indeed the case for $D = 2, H = 1$ in the argument given here. However, the social interaction picture, constrained by transportation costs across the city, is more fundamental because it can be fulfilled by individuals, appeals directly to function and social dynamics and does not require global optimization.

$H < 1$ corresponds to a trajectory with a volume less than linear and is in practice a series of separate spatial clusters. This means that an individual cannot reach the entire city by himself, though the city may still stay connected via a chain of local interactions. While a city can exist as such, cities would potentially become more and more disconnected as they grow, requiring a larger number of overlapping zones and interpersonal contacts to be available to each citizen. In this regime a city would then behave more like a series of separate interacting communities rather than a whole mixing population, a characteristic that is often used to define the absence of a *bona fide* functional city. Note that in the limit $H \to 0$, the exponent $\delta = \frac{H}{D(D+H)} \to 0$, and urban agglomeration effects (superlinearity of socioeconomic outputs and sublinearity of infrastructure) altogether vanish. Given available data, which provides typically only total administrative unit area for a city or metropolitan area, this $H \lesssim 1$ regime seems to be sometimes observed, see discussion below and Table S3.

Conversely, $H > 1$ means that the length of trajectories scales faster than a linear volume, and in particular for $H = 2$ they would scale as an area (and for $H = 3$ as a 3D volume). Because cities are approximately two dimensional we may expect $H \leq 2$ to be an absolute upper bound, which leads necessarily to $\alpha \geq 1/2$ and $\delta \leq 1/4$. It is important to stress that although individuals may explore the city in a way that is area filling locally, this does not imply that $H = 2$ in general. This is because the characteristic length is measured in terms of the area of the city, and consequently $H = 2$ would mean that they would have to cover the entire land area over a given time period. This seems counterfactual, certainly for large cities. For all these reasons, while I leave H as a parameter in the main text, I expect that it would naturally be of

order $H \simeq 1$, with $\alpha = \frac{D}{D+H} \simeq 2/3$ and $\delta = \frac{H}{D(D+H)} \simeq 1/6$, see also main text.

Infrastructure Networks' Length is Area Filling

I have assumed in the main text the property that networks of infrastructure fill the area of the city. This assumption is implicit in the principle that infrastructure networks grow in a decentralized way in order to connect each addition of a new inhabitant. This assumption means more explicitly that any occupied land area (as residence, business or any other use) can be reached by people, goods and information traveling over infrastructure networks. The technology involved in these networks varies enormously with level of urban development but I assume here that the geometry of the networks does not. Figure 2A (main text) illustrates this situation for a regular grid. In this case the total length of the network can be derived easily, see Figure 2A (main text), as

$$A = L^2 = (n_b l)^2; \qquad L_n = 2(n_b + 1)L = 2(n_b + 1)n_b l = \frac{2}{l}A + 2\sqrt{A} \underset{n_b >> 1}{\sim} A, \qquad \text{(S8)}$$

where l is the average block length (the minimum separation along the network), n_b is the (linear) number of blocks across the city, and $L = n_b l$. The factor of 2, in the first term of L_n above, accounts for vertical plus horizontal network segments, and the factor of $n_b + 1$ counts the number of segments across the city, including one at the edge, each with length $L = n_b l$. The factor of n_b^2 that results is then identified with the area A, up to a multiplicative constant. For networks that are not, on the average, square grids the constants multiplying the factors of area, A, will differ, but not the space filling character of the network, expressed as $L_n \simeq A/l$.

Boundary Conditions and Scaling of Currents

Here I show more explicitly the effect of the choice of boundary conditions on network model variables and the introduction of certain city size independent individual and network properties. This choice is important because it sets the scaling behavior of energy dissipation in the network

due to transportation processes. I have assumed that the width of the smallest network units, s_*, is a constant, independent of city size. Although seemingly an abstract assumption this means in practice something quite intuitive, that house doors, water faucets and electrical outlets, for example, each have a common cross section in all cities that does not vary with city population size. This means that I can write the scaling of width across network levels as

$$s_i = s_* b^{(\delta-1)(i-h)}, \tag{S9}$$

which implies that the width is largest at the highest level ($i = 0$: root, "highways") $s_0 = s_* b^{h(1-\delta)}$, since $b > 1$ and $\delta << 1$. In addition recall that $N_i = bN_{i-1}$, $N = N_h = b^h$ and that it follows from the conservation of current, J_i, that

$$J_i = s_i \rho_i v_i N_i = s_{i-1} \rho_{i-1} v_{i-1} N_{i-1} = J_{i-1}, \qquad \forall_{i=1}^{h}. \tag{S10}$$

This condition may apply only statistically (*38*) for a network that is not a (balanced) tree, as for example, would happen in a semi-lattice (*26*), where branches at the same level are connected, or upper branches can converge on the same lower site. This condition leads to the scaling relation for the current density

$$\rho_i v_i = b^{-\delta} \rho_{i-1} v_{i-1}. \tag{S11}$$

This relationship is not fully specified until I prescribe its boundary conditions. I can place a limit on the current density at the root $\rho_0 v_0 = \text{const}$, which leads to $\rho_i v_i = b^{-\delta i} \rho_0 v_0$, or at the smallest branches $\rho_h v_h = \rho_* v_*$, which leads alternatively to $\rho_i v_i = b^{\delta(h-i)} \rho_* v_*$. These conditions result in the forms for the total current at each level

$$J_i = s_i \rho_i v_i N_i = s_* \rho_0 v_0 b^{h(1-\delta)}, \tag{S12}$$

or

$$J_i = s_* \rho_* v_* b^h, \tag{S13}$$

respectively. Both these forms are independent of level i, a necessary consequence of total current conservation, but they scale with population size in different ways. Specifically, given a boundary condition at the root obtains $J_i = J = s_* \rho_0 v_0 N^{1-\delta}$, while for the boundary condition at the leaves this leads to $J = J_0 N$, with $J_0 = s_* \rho_* v_*$. Note that the latter is the expected current for a population of individuals, in terms of their intrinsic "individual needs", and is therefore the natural boundary condition. It means, in intuitive terms, that the flow of people through doorways in their homes is similar across cities of different sizes and that the consumption of water, electricity, etc, per capita in households is an invariant of city size, as observed (*12*). Thus, the differences between cities arise at larger scales, where social interactions are more common and population-wide constraints apply. Thus, life at home in cities of any size remains in many ways the same; it is only in public interaction spaces that the more urban character of larger cities manifests itself.

Dissipation on Infrastructure Networks

There are many dissipative processes (costs) that can take place in a city and that can lead to situations in which increasing social interactions and their products may be more than overcome by their associated costs. In the main text I assume the the resistance at each level of the network is that of all branches taken in parallel (c.f. (*4*)), that is

$$R_i = \left[\sum_{i=1}^{N_i} \frac{1}{r_i} \right]^{-1} = \frac{r_i}{N_i}, \tag{S14}$$

as usual, if all branches have the same resistance r_i. The resistance of each branch is a purely geometric property of the network times a resistance, r, per unit length and transverse area,

$$r_i = r \frac{l_i}{s_i} = r \frac{a}{l s_*} b^{(\alpha-\delta)i + (\delta-1)h}, \tag{S15}$$

which increases with level i, and therefore is larger in the smallest branches than at the root. From (S14) this leads to

$$R_i = r\frac{a}{ls_*}b^{-(1-\alpha+\delta)i+(\delta-1)h}, \tag{S16}$$

which *decreases* with i and is therefore larger at the root (highways) than at the leaves (narrow local paths). This is a direct result of the assumed parallelism of the branches at each level. If they are not strictly operating in parallel then the total resistance will decrease less slowly from the root to the leaves of the network, and be larger in total, leading to higher dissipation than estimated here. We can put the conditions on the current and resistance together to obtain the total power dissipated, W, as

$$W_i = R_i J^2, \tag{S17}$$

$$W = \sum_{i=1}^{h} W_i = J^2 \sum_{i=1}^{h} R_i = rJ^2\frac{a}{ls_*}b^{(\delta-1)h}\frac{1-b^{-(1-\alpha+\delta)(h+1)}}{1-b^{-1+\alpha-\delta}} = W_0 N^{1+\delta}, \tag{S18}$$

which scales superlinearly, with an exponent $1+\delta \simeq 7/6$ ($D=2, H=1$). The pre-factor in Eq. (S18) is $W_0 \simeq \frac{arJ_0^2}{ls_*(1-b^{-1+\alpha-\delta})}$. We see that the dissipative behavior of the network is set by the current squared, J^2, multiplied by the resistance at the root, $R_0 = \frac{ar}{ls_*}N^{\delta-1}$. The current, in turn, is set by conditions at the smallest branches, that is, by the fundamental properties of people and their behavior. Thus, the main overall contribution to these dissipative processes results from people, energy, information, etc, being channeled through a network with many levels, and of the constraints that occur at its largest scales. Remarkably, this result ties together the most microscopic needs and behaviors of individuals anywhere to the most macroscopic aspects of the urban infrastructure.

Another way to see this is to rearrange terms in Eq. (S18) to write it as

$$W = r'\left(\frac{a}{l}\right)^2\frac{J^2}{A_n} \tag{S19}$$

where $r' = \frac{r}{(1-b^{\alpha-\delta-1})(1-b^{\alpha+\delta-1})}$. This shows that that the dissipation term can be made smaller by increasing the infrastructure network's total volume, A_n. In contrast, as we have seen above, making A_n smaller increases the social outputs of cities. Thus, we may expect an equilibrium between the detailed consequences of these two effects that leads to an optimal allocation of infrastructure to social interactions as a function of population size (and level of technology).

Global Optimization

Here I show that the principles discussed in the main text can be formulated in terms of a constrained optimization problem, where each individual maximizes the outcome of his/her interactions minus costs, subject to the general infrastructural and size constraints posed by the city, and where city infrastructure can be managed so as to maximize collective welfare. I write the objective function, $\mathcal{L}$, for this problem as

$$\mathcal{L} = Y - W + \lambda_1 \left(\epsilon A^{H/D} - GN/A\right) + \lambda_2 \left(A_n - cNd\right), \tag{S20}$$

where $c = A_0 a^{-1/D}$ is a constant in N that follows from Eqs. 2 and 4, $d = (A/N)^{1/D}$ (see main text) and λ_1, λ_2 are Lagrange multipliers. From their point of view, individuals can structure their interactions in space and time so as to maximize the benefit of being in the city, while minimizing costs. This is expressed primarily in terms of the factors that enter G. In turn, city authorities should provide organizations (such as police, which affect social interaction modes) and infrastructure so that general urban socioeconomic benefits are maximized. This can be expressed in terms of the variation of A_n (and of the factors that make it). Varying (S20) relative to A and A_n leads to

$$Y(N) = G\frac{N^2}{A_n(N)}, \qquad W(N) = r'\left(\frac{a}{l}\right)^2 \frac{J^2(N)}{A_n(N)}, \tag{S21}$$

that is, it imposes the dependences in N of A and A_n discussed in the main text and their consequences for social outputs and network dissipation.

Now observe that the problem of matching the sum total of social interactions to costs has two solutions in terms of values of G, specifically

$$G \equiv G_{\min} = 0, \qquad \text{or} \qquad G \equiv G_{\max} = \left[\frac{(\epsilon l)^{2\alpha}}{r' J_0^2} l^{2(1-\alpha)}\right]^{\frac{1}{2\alpha-1}}. \tag{S22}$$

The first solution, at $G = 0$, means that for a city to exist it needs to have some level of net positive social interactions, $G > 0$. The second solution is the point at which network dissipation costs overwhelm the social benefits of the city, beyond which the city becomes too expensive to exist as a whole and may break up into disconnected areas. In between these two extremes, there is a special value of the coupling $G = G^*$ for which the balance is positive and largest. We can determine this point by taking the variation of the net benefits $\mathcal{L}$ relative to G, (recall that $a = (G/\epsilon)^\alpha$), to obtain

$$\frac{d\mathcal{L}}{dG} = \left[(1-\alpha) - \alpha \frac{r' J_0^2}{G} \left(\frac{a}{l}\right)^2\right] \frac{N^2}{A_n(N)} = 0, \tag{S23}$$

which results in the solution

$$G = G^* = \left[\frac{1-\alpha}{\alpha}\right]^{\frac{1}{2\alpha-1}} G_{\max} \leq G_{\max}. \tag{S24}$$

This condition implies that there is an optimal G to which any city should converge in order to maximize its difference between net social output and associated dissipation costs. Note that the city can only exist if social outputs are larger than dissipation and that, starting with small $G > 0$, it pays to increase the coupling for a while. However, increasing it beyond $G > G^*$ leads to dissipation rising faster than social outputs, reducing the net difference between the two and ultimately canceling them altogether.

Finally we can rewrite $\mathcal{L}$ at G^* as

$$\mathcal{L} = Y - W = \frac{2\alpha - 1}{\alpha} G^* \frac{N^2}{A_n}. \tag{S25}$$

We see, therefore, that the optimization that is achieved in the city is open-ended relative to population size, N, as long as both individual choices and infrastructure can be mutually adapted

to (close to) their optimal values. This emphasizes the interplay between individual and social behavior, which constitutes the necessary condition for the city to exist, and the role of infrastructure and policy in creating the conditions that promote the benefits and reduce the costs of human social behavior.

In practice, these conditions predict that the most likely value of G from a sample of many cities should correspond to G^*. Therefore, we can use an estimate of the mode of the distribution of G (Fig. 1B inset) to obtain G^*, shown as the solid yellow line. Thus, monitoring relative changes in Y (e.g. via GDP) versus those in W (the dissipative costs of transportation) allows cities to judge how close to their optimal net output they are. More importantly, it also provides a practical and quantitative basis for iterative city management and policy as well as for benchmarking urban areas relative to each other, see Fig. 1B (inset).

Empirical Support for Urban Scaling Laws Predicted by Theory

Table S3 (and Table 1) summarizes the empirical literature on estimates of scaling relations for cities including a few new ones, introduced here. I now discuss some of the current evidence for the ranges of exponents given in Tables 1 and S3 and present a few additional examples that emphasize the consistency of scaling exponents with data from many different urban systems and over time. I also discuss issues related to currently available data for specific urban indicators and directions for future empirical research.

There have been over 50 years of research characterizing many urban properties in terms of allometric, or power law, relations, especially in geography, sociology and urban economics. The use of these types of scale-invariant relations is dictated by the general, but often unstated, assumption that human settlement properties, from the smaller towns to the largest cities, vary continuously and that there is no particular population or length scale at which they change radically. This is supported by a vast body of empirical evidence, only a fraction of which is

discussed below.

Urban Scaling of Land Area: There have been many attempts over the last 50 years at characterizing the scaling relation between the land area of cities, A, and their population, N. Many of the early characterizations of this scaling relation in the geography literature use definitions of cities in terms of sets of administrative units (*39*) (e.g. counties, municipalities), which leads to several potential biases. Also, more modern methods relying on remote sensing measure built area (impervious surfaces), which more closely tracks the infrastructure network, especially in larger cities, thus providing a good measure of A_n, but not necessarily of A, the circumscribing area, see Table S3. In addition, some of these studies, especially early analyses of satellite data (*40, 41*), assumed that cities would take an approximately circular form, and analyzed the scaling of the corresponding radius, see also (*42*). These methodological choices introduce uncertainly in the measurement of the area of cities and are further compounded by changes in settlement spatial profiles over time that occur as a result of socioeconomic growth and technological change in transportation, as discussed in the main text. These effects must be carefully considered in order to compare statistical analyses of data.

The first quantitative analyses of the relation between A and N, to the best of my knowledge, go back to the late 1950's and early 1960s. Stewart and Warntz (*39*) measured areas associated to cites (political units) in the US in 1940 and England and Wales in 1951 and found $\alpha \sim 0.75$. Nordbeck (*18*), using 1960 and 1965 data, found that for about 1,800 urbanized areas in Sweden $\alpha \simeq 2/3$, which is also consistent with US metropolitan statistical areas in recent years (*42*). He discusses different definitions of cities, but uses the Swedish *tätort* (urbanized area) in his empirical analysis, which is a built up area connected by infrastructure in analogy to the principles and assumptions made in the main text. Nordbeck's work (*18*) is important because, in addition to this empirical analysis, it develops the first theoretical argument for $\alpha = 2/3$ and discusses the nature of the settlements that show the greatest variation

from scaling, see also (*43*). Specifically, he discusses the particular functional characteristics of settlements, especially among small towns, that can lead to deviations from scaling, from dormitory towns and railroad villages, to fishing communities and resorts. But most importantly, Nordbeck develops a theoretical argument for cities in terms of general allometry, in analogy to river networks and biological organisms. He argues that the the settlement profile of cities is heterogeneous in space and that, as a consequence, total population must scale with a dimension larger than that of land area and has a most natural dimension of a physical volume ($D = 3$). This implies $N \sim A^{3/2}$, or $A(N) \sim N^{2/3}$, as he observes (*18*). The theoretical framework developed in the main text derives this result in the particular case of $D = 2, H = 1$. But, importantly, the theory developed here shows how this scaling exponent can be obtained through the consideration of more fundamental urban function rather than form, specifically via the requirement of population mixing under advantages of agglomeration in terms of social interactions, subject to transportation costs.

After Nordbeck, several other studies found values of α in the range $2/3 \leq \alpha < 1$, see e.g. (*39–42, 44–47*). The larger observed variability in the exponent of this scaling relation relative to others may be the result of the adoption of different (inconsistent) definitions of city, limited ranges of scales, different measurement methods, etc (*48*). The range for α given in Table 1 (see Table S3) is a synthesis of these results and reflects this larger variance. It would be desirable in the future to develop a more consistent approach to measuring this scaling relation, hopefully motivated by the theoretical framework developed here.

Urban Scaling of Paved Areas: Several of the more modern empirical studies aiming at capturing the spatial extent of cities use satellite imagery as a means to measure the built area of settlements. These data allow, in principle, for simultaneous and consistent measurements over many cities. Pioneering studies by Batty, Longley and collaborators in the early 1990s (*49*), analyzing specific regions of Britain (such as Norfolk) found no (strong) agglomeration effects

in this regional context, and developed null models to account for such behavior. However, these techniques have since developed further and been carefully calibrated and refined over the last 20 years, see e.g. discussion in (*16*). As a consequence remote measurements of built area have probably become reliable in the last few years thanks to new satellite imagery and more thorough comparative analyses. When applied to large cities they measure primarily the area of paved infrastructure, thus providing a measure of A_n, not A. A variety of recent results for large cities in Europe (*50*), China (*51*) and over 3,600 cities over 100,000 people worldwide (*16*) establish the scaling $A_n \sim N^{\nu}$, with $\nu \sim 5/6$ (see Table S3), in general agreement with the theoretical arguments developed in the main text. The value of the exponent ν shown in Table 1 is a synthesis of direct measurements, such as those shown in Fig. 1A (main text), and of these recent remote sensing estimates, see Table S3. More systematic measurements of A and A_n and corresponding population sizes, would be desirable and new remote sensing datasets, properly calibrated and expanded to smaller settlements, may provide the best candidate empirical approach to this end (*16*).

Urban Scaling of Social Outputs: Perhaps an even richer literature addresses the effects of city population size on urban socioeconomic quantities. General qualitative arguments for the advantages of cities in human social life are very old and date back to at least Aristotle (*52*) in his *Politics*, where he discusses the greater scope of human sociality in cities (polis) when compared to animal societies. In terms of inspiring economic theory, the work of Alfred Marshall about industrial districts (*23*) and, more recently, of Jane Jacobs (*22*) and Allan Pred (*53*), suggested general qualitative mechanisms for why larger cities can generate increases in innovation and economic production rates and inspired many subsequent empirical and theoretical studies. The main arguments about these dynamical (functional) advantages of larger cities rely on aspects of their internal socioeconomic structure, which allow larger cities to provide new and better services in the context of a (national or even international) system of cities (*22*), (*53*). These ideas,

which together with concepts from location and central place theory (*54, 55*), constitute the basis for modern economic geography (*29*), require progress in the current quantitative understanding of urban social and infrastructural networks as the basis for predicting agglomeration effects. This is the objective of the theoretical framework developed in the main text.

Historically, it was only in the 1970s, to my knowledge, that direct empirical analyses of socioeconomic output rates across cities were first carried out. Sveikauskas (*9*) provided one of the first such measurements, dealing with the scaling of value-added in several manufacturing sectors in US metropolitan areas with population, controlling for the education level of the workforce. In this spirit, many subsequent studies followed and attempted to control for other factors that co-vary with city population size, such as types of economic sector or measures of human capital. It is important to emphasize that these factors vary over time and with the level of socioeconomic development, while scaling exponents do not, see Fig. S3. For example, much of the wealth of cities in China at present, see Fig. S1, depends on manufacturing, while the wealth of cities in Germany, Japan or the USA, Figs S1-3, rely on other more "high-tech" sectors, such as media, finance, and more advanced technology, with different compositions in different nations. The disproportionate concentration of high value-added sectors and human capital in larger cities provides only a circular argument for the higher productivity of larger cities, a point already clearly made by Jacobs (*22*) and Pred (*53*) in their foundational work on the economies of cities. Thus, these effects must be explained together from a more fundamental dynamics that allows both increases in economic productivity and in human capital to arise in the first place and reinforce each other to realize their productive potential (*22*,*23*), (*53*). A central objective of the theoretical framework of the main text is to provide a basis for the understanding of such joint dynamics in terms of evolving structures of urban social networks.

The most direct measure of the size of urban economies is the Gross Domestic Product (GDP) of metropolitan areas. The US Bureau of Economic Analysis has been publishing data

on Metropolitan GDP since 2001 only, and of components of personal income (including wages) since 1969, see Fig S3. Metropolitan GDP is derived as the sum of the GDP originating in all the industries in the metropolitan area, at some level of industrial classification resolution. Both these measures (GDP, wages) rely on data collection from a variety of state and federal sources and their construction is complex but consistent, see `http://www.bea.gov/regional/methods.cfm`. For other nations these procedures tend to be analogous and some detail is available in the original sources' online materials, see Materials and Methods above.

Analyses of other quantities mediated through social interactions, such as crime, the incidence of contagious diseases, innovation, etc provide opportunities to measure the effects of cities in accelerating human social contact rates. Generally, in modern societies, levels of person-on-person crime (such as homicides) increase superlinearly with city size (*10*,*12*,*25*). However, this may not always have been the case as it is the net result of superlinear opportunities for violence and social measures to combat crime (metropolitan police), which historically are first developed systematically in larger cities (*1*), (*56*). It has been argued for example that in Medieval times, cities were safer than the countryside (and both were very violent by modern standards) (*1*), and also that pre-1940s larger cities in the USA were safer (*57*) than smaller places, but quantitative data in support of these statements are weak. At present, urban systems that present higher levels of violence, especially in Latin America and parts of Africa, also suffer from severe issues affecting the availability and reliability of reported data (*25*). Nevertheless, where data are systematically collected exponents are in the expected ranges (*25*).

These reporting biases also affect measurements of public health data, especially dealing with the incidence of contagious diseases. In developed nations, access to modern public health in large cities has all but stemmed the worst impact of most contagious diseases. An exception may be HIV/AIDS in the early years of the epidemic (*12*) (which showed superlinear rates of incidence with city size), as no crowd immunity effect or clinical inter-

vention could then reduce death rates. The effect of antiretroviral treatment for the disease has sharply reduced new cases of the disease (as well as deaths) from the date of its introduction in the US in 1996. But because this introduction seems to not have suffered from substantial city size biases (at least among large cities for which there is data), it has not affected scaling exponents significantly. Data from the US Centers for Disease Control (CDC) for the earliest years of the epidemic are sparser and concentrated on higher risk population segments, see `http://www.cdc.gov/hiv/topics/surveillance/resources/reports/past.htm#supplemental`, and the specific urban areas where the epidemic first took root, such as San Francisco and New York City. Later years provide a more general picture of urban contacts leading to the spread of the disease. It would be interesting to perform more thorough analyses of the epidemic's history, with these events and data limitations in mind.

Other contagious diseases, such as diarrheal diseases, have historically contributed to much urban mortality, especially among children. Where data are available, the incidence of these diseases is clearly correlated to urbanization, including in 19th century Britain (*58*). At present, where these diseases remain a major burden, data are usually lacking or are unreliable. This situation may soon change, due to new measurement opportunities in developing nations resulting from portable device technologies. These opportunities will provide important tests and applications of the theoretical framework developed here in environments where understanding and managing urbanization are most critical.

Urban Scaling of Invention and Innovation: While very important for growth and development, innovation is difficult to measure unambiguously. Patents (*11*) have provided a proxy for technological innovation, as have employment in 'creative' sectors (*59*). Patents tend to show scaling exponents slightly larger than most other socioeconomic rates, in the range $\beta = 1.2 - 1.3$, possibly because they rely on interactions between individuals that are already dis-

proportionately present in larger cities. For example, the number of supercreative[1] professionals (*59*) scales with city size in the US with an exponent $\beta \simeq 1.15$ (*11*).

The exponent range for socioeconomic rates in Table 1 are obtained from references (*9*,*10*,*11*,*12*), (*43*), without controlling for co-varying factors, see Table S3.

Direct Measurement of Scaling of Social Interactions: Direct measurements of human social interactions are difficult to obtain because of issues of privacy and because of the vast numbers of people involved in each city and the necessity to cover cities of different sizes. Pioneering studies in sociology (*60,61*), based on survey data, established a few general trends about strong links and how they vary with types of people, such as age or socioeconomic status. These findings suggest an increase in social connectivity with city size, as more people of working age and of higher socioeconomic status manifest on average more connections, but their methods and scope are not entirely conclusive in this respect. More recently, the advent of large scale cell phone data allows us more comprehensive measurements of social networks, though with a different set of caveats. A recent study of two different European nations (*21*), Portugal and the UK, establishes the superlinear behavior of various measures of social interactions with exponents in general agreement with theoretical expectations, see Table S3. This is the range shown in Table 1.

Urban Scaling of Land Rents: An estimate for land rents across the city, given in Tables 1 and S3, follows from considering the total income $\sim N^{1+\delta}$ divided by the total land $\sim N^{\alpha}$, which results in average land rents (measured in units of money per unit area and unit time) scaling with an exponent $P_L \sim N^{1-\alpha+\delta}$, $1-\alpha+\delta \simeq 1/2$ in $D=2, H=1$. Thus, land rents scale faster with population than incomes or wages. This is offset in part, by smaller per capita use of land, achieved primarily by increasing the floor area of buildings relative to their land footprint

[1]Supercreative professionals (see (*59*), pages 327-329) are defined as individuals working in "Computer and Mathematical, Architecture and Engineering, Life, Physical and Social Sciences, Education Training and Library, Arts, Design, Entertainment, Sports and Media" occupations. These professions are defined by the Standard Occupation Classification System of the US Bureau of Labor Statistics.

by building taller multi-floor units.

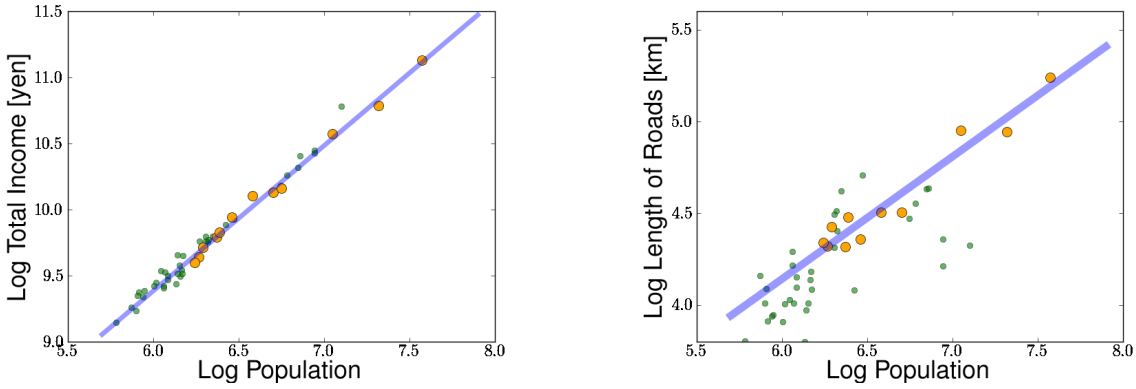
Data on land rents across cities are fairly sparse. Ranges given in Table 1 are obtained by adapting the results from (*24*) on housing values as a proxy. Note that in their Figure 3 these authors estimate the scaling of personal income per capita on median house value by metropolitan area. They find an exponent of 0.34 ± 0.02. Thus, with income per capita scaling with exponent δ, we obtain that $P_L \sim N^{\delta_L}$, with $\delta_L = \frac{\delta}{0.34} = 2.94\delta = 1-\alpha+\delta$. In the estimates given in Table 1 I used $\delta = 1/6$, but a slightly lower value of this exponent (often associated with income, but not wages or GDP, because of national transfers) will naturally result in a lower estimate of δ_L.

New Evidence for Predicted Scaling Relations for Different Nations and Time: Figs. S1-3 show how the scaling of superlinear socioeconomic rates is a property of many urban systems worldwide, across several continents (Americas, Asia, Europe) and levels of development (e.g. China or the USA). Fig. S1 also shows the sublinear scaling of infrastructural quantities, see also (*12*), in relation to corresponding socioeconomic rates for the metropolitan areas of Japan. Finally, Fig. S3 show the extraordinary consistency of superlinear scaling of socioeconomic rates (wages) in US Metropolitan Statistical Areas over the last 40 years, see also (*43*). This confirms the theoretical prediction that scaling exponents are independent of time, and levels of socioeconomic development, which vary considerably over this period; see Fig. S3d.

All this evidence, across time and nations all over the world, establishes some of the general properties of cities that can explain their universal role in socioeconomic development of human societies (*1*,*9*,*22*). Current data gaps, uncertainties and potential biases also stress the need for theory that can establish key quantities to measure and provide quantitative hypothesis for their magnitude and behavior. The framework developed here seeks to explain and provide predictions for urban data anywhere, through an interdisciplinary quantitative synthesis from geography, sociology, urban economics, planning and complex systems. It establishes the

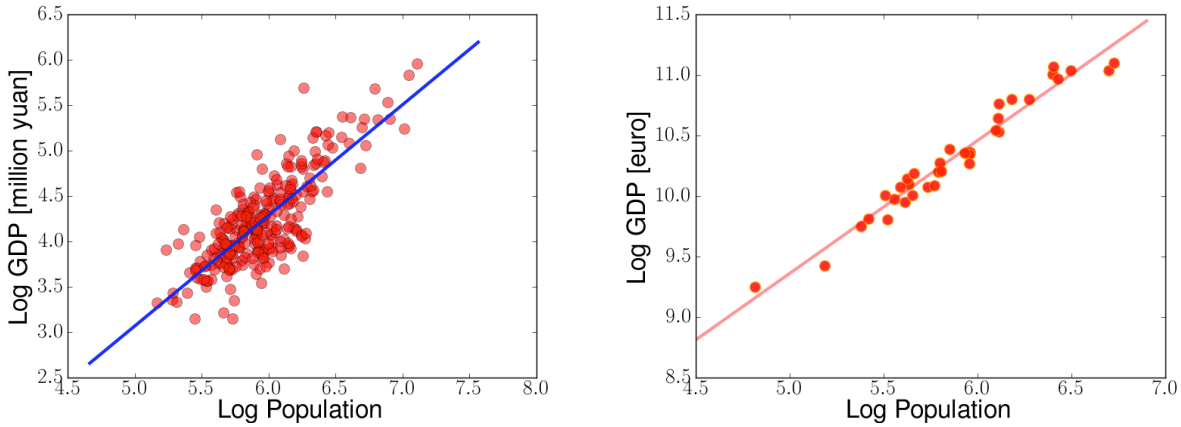
fundamental nature of cities in terms of network theories of people and infrastructure.

Supplementary Figures



(a) (b)

Figure S1: Income and infrastructure for Japanese cities in 2005. (a) Scaling of the income in Japanese Metropolitan Areas (MAs) (orange circles) and underlying prefectures (green dots). Observed scaling with exponent $\beta = 1.12$ (95% confidence interval $[1.07, 1.17]$, $R^2 = 0.99$) is shown as the solid blue line. (b) Scaling of the the length (not area) L_n of roads in Japanese MAs (orange circles) and prefectures (green dots). Observed exponent (solid blue line) is $\beta = 0.67$ (95% confidence interval $[0.55, 0.78]$, $R^2 = 0.94$). Both exponents are in close agreement with the theoretical expectations developed in the main text.



(a) (b)

Figure S2: Scaling of Gross Domestic Product (GDP) for cities in China and Germany. (a) Scaling of the GDP of prefectural cities (urban districts) in China in 2005. Observed scaling with exponent $\beta = 1.22$ (95% confidence interval $[1.11, 1.33]$, $R^2 = 0.65$) is shown as the solid line. Urban population is measured through official residence permits. There is substantial evidence that population counting in this way underestimates the resident population of large cities by as much as 5 million in Beijing and Shanghai. Such adjustments would bring the estimated exponent down slightly. (b) Scaling of the GDP of German Larger Urban Zones (functional cities defined by European Union Statistics) in 2004. Observed exponent (solid red line) is $\beta = 1.10$ (95% confidence interval $[1.02, 1.18]$, $R^2 = 0.96$). The estimate of the exponent is brought down slightly by the two largest cities, the Ruhr Valley and Berlin, respectively. In their absence the estimated scaling exponent agrees perfectly with the simplest theoretical expectation of $\beta = 7/6$. The Ruhr Valley is an industrial area composed of several distinctly recognizable cities. Berlin is a special large city having been largely destroyed in World War II and still experiencing, at this time, the integration of its Western and Eastern parts, in the aftermath of German reunification. Its current population of about 3.5 million remains substantially smaller than at its height (in 1939 it was estimated at 4.3 million.)

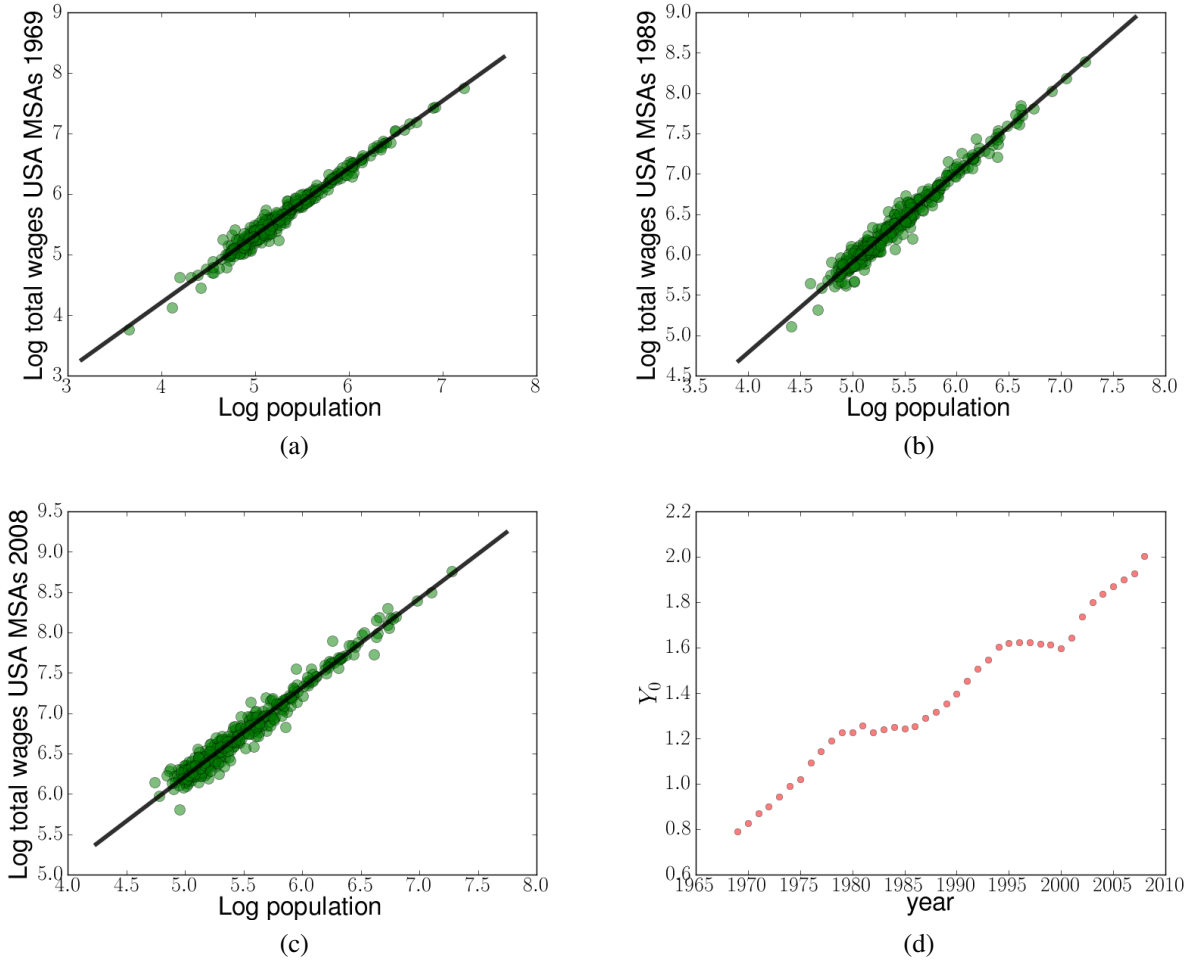


Figure S3: Consistency of scaling relations for socioeconomic outputs over 40 years. Scaling of total wages (in thousands of US$) in US Metropolitan Statistical Areas (a) in 1969, scaling exponent $\beta = 1.11$ (95% confidence interval $[1.09, 1.13]$, $R^2 = 0.95$) (b) in 1989, $\beta = 1.12$ (95% confidence interval $[1.10, 1.14]$, $R^2 = 0.94$), and (c) in 2008, $\beta = 1.10$ (95% confidence interval $[1.08, 1.13]$, $R^2 = 0.93$). Throughout this 40-year period exponents stayed remarkably consistent, within each other's 95% confidence intervals, even as population and baseline wages, Y_0, grows considerably (d), and the economic base of cities changes profoundly, undergoing several periods of recession and expansion.

Supplementary Tables

Input Variable	Symbol
dimension of space	D
fractal dimension for travel paths	H
transportation cost per unit length travelled	ϵ
interaction area per person (cross section)	a_0
average travelled distance per person (path length)	ℓ
interaction strength for social quantity k	g_k
smallest network cross section	s_*
current density in smallest network segments	ρ_*
carrier velocity in smallest network segments	v_*
resistance per unit length of network	r

Derived Urban Parameter	Symbol
average social interaction strength	$\bar{g} = \sum_k p(k) g_k$
coupling	$G = \bar{g} a_0 \ell$
agglomeration exponent	$\delta = \frac{H}{D(D+H)}$
current per capita on networks	$J_0 = s_* \rho_* v_*$

Table S1: Fundamental urban parameters and their significance.

Scaling Relation	Pre-factor	Exponent
land area	$a = \left(\frac{G}{\epsilon}\right)^{\alpha}$	$\alpha = \frac{D}{D+H}$
network volume (area)	$A_0 = \frac{s_* a}{l} \frac{1}{1-b^{\alpha+\delta-1}}$	$\nu = 1 - \delta$
network length	$L_0 = a/l$	$\lambda = \alpha$
social interactions and socioeconomic rates	$Y_0 = G/A_0$	$\beta = 1 + \delta$
power dissipation on networks	$W_0 = \frac{a r J_0^2}{s_* l (1-b^{-1+\alpha-\delta})}$	$\omega = 1 + \delta$
land rents	$P_0 = Y_0/a$	$\delta_L = 1 - \alpha + \delta$

Table S2: Summary of the dependence of scaling parameters on input variables (Tables 1 and S1). Note that exponents are only dependent on the the dimensionless parameters H, D, and are in general independent of network parameters or of details of individual behavior. In this sense we may expect exponents to be largely invariant in time, population size or levels of socioeconomic development. Nevertheless, H is a means of measuring how connected (inclusive) a city is and may change slowly over time. The prefactors of the scaling relations depend on the remaining input parameters and do change over time (e.g., see Fig. S3d), reflecting socioeconomic development and changes in the properties of infrastructure and individual behavior.

Scaling Relation	Exponent	Error	Observations	Region/Nation	Urban Unit	Year	Reference
Land area							
administrative	$\alpha = 0.75$	NR	412	USA	Cities (political)	1940	(*39*)
administrative	$\alpha = 0.75$	$R^2 = 0.87$	157	England and Wales	Cities (political)	1951	(*39*)
urbanized	$\alpha = 0.66$	[0.65,0.67]	1,800	Sweden	Tätort (urban area)	1960	(*18*)
urbanized	$\alpha = 0.65$	[0.64,0.66]	1,800	Sweden	Tätort (urban area)	1965	(*18*)
urbanized	$\alpha = 0.63$	0.62-0.64	329	USA	MSA	1980-2000	(*42*)
developed	$\alpha = 0.57$	0.56-0.59	329	USA	MSA	1992, 2000	(*42*)
light emissions	$\alpha = 0.65$	$R^2 = 0.62$	4,851	USA	Night-light clusters	1992	(*47*)
Average land area	$\alpha = 0.67$	[0.56,0.75]					
built area	$\alpha = 0.78$	NR	89	USA	Cities (political)	1960	(*44*)
built area, radial	$\alpha = 0.88$	NR	368†	Michigan, USA	Cities (political)	1969	(*40*)
built area	$\alpha = 0.96$	[0.89,1.04]	70	Norfolk, UK	Settlements	1981	(*41*)
built area, radial	$\alpha = 0.87$	[0.75,0.99]	70	Norfolk, UK	Settlements	1981	(*41*)
built area	$\alpha = 0.87$	NR	51	Ontario, Canada	Urban Areas	1966	(*46*)
Average land area*	$\alpha = 0.75$	[0.56,1.04]					
Network area (or volume)							
impervious surfaces	$\nu = 0.85$	0.84-0.86	3,629	World	Cities$> 100,000$	2000	(*16*)
impervious surfaces	$\nu = 0.86$	$R^2 = 0.74$	119	EU	Agglomerations$> 200,000$	1990	(*50*)
built area	$\nu = 0.82$	$R^2 = 0.84$	660	China	Urban Areas	2005	(*51*)
area of roads	$\nu = 0.85$	[0.81,0.89]	451	USA	MSA	2006	Fig. 1A
area of roads	$\nu = 0.83$	[0.74,0.92]	29	Germany	LUZ	2002	(*12*)
Average network volume	$\nu = 0.84$	[0.74,0.92]					
Network length							
length of pipes	$\lambda = 0.67$	[0.55,0.78]	12	Japan	MA	2005	Fig. S1
Socioeconomic rates							
GDP	$\beta = 1.13$	[1.11,1.15]	363	USA	MSA	2006	Fig.1B,(*12*)
GDP	$\beta = 1.22$	[1.11,1.33]	273	China	Prefectural Cities	2005	Fig. S2A
GDP	$\beta = 1.10$	[1.01,1.18]	35	Germany	LUZ	2004	Fig. S2B
income	$\beta = 1.12$	[1.07,1.17]	12	Japan	MA	2005	Fig. S1A
wages	$\beta = 1.12$	[1.07,1.17]	363	USA	MSA	1969-2009	Fig. S3
violent crime	$\beta = 1.16$	[1.11,1.19]	287	USA	MSA	2003	(*12*)
violent crime	$\beta = 1.20$	[1.07,1.33]	12	Japan	MA	2008	(*62*)
violent crime	$\beta = 1.20$	[1.15,1.25]	27; 5,570	Brazil	MA; Municipios	2003-07	(*25*), (*62*)
new AIDS cases	$\beta = 1.23$	[1.17,1.29]	93	USA	MSA	2002-3	(*12*)
new patents	$\beta = 1.27$	[1.22,1.32]	331	USA	MSA	1980-2001	(*11, 12*)
supercreative jobs	$\beta = 1.15$	[1.13,1.17]	331	USA	MSA	1999-2001	(*11*)
R&D employment	$\beta = 1.19$	[1.12,1.26]	227-278	USA	MSA	1987-2002	(*11*)
Average socioeconomic rates	$\beta = 1.17$	[1.01,1.33]					
Social interactions							
cell phones	$\beta = 1.12$	[1.00,1.25]	415	Portugal	Cities, LUZ, Municipality	2006-7	(*21*)
land lines	$\beta = 1.12$	[1.05,1.17]	24	UK	Cities	2005	(*21*)
Average social interactions	$\beta = 1.12$	[1.00,1.25]					
Power dissipation							
electrical	$\omega = 1.11$	[1.05, 1.17]	380	Germany	Cities	2002	(*12*)
Average land rents							
median house value	$\delta_L = 0.49$	[0.46,0.52]	363	USA	MSA	2006	(*24*)

NR=not reported. Error, in order of availability from the source, is given by: 95% confidence intervals (square brackets), ranges, or R^2 values.
Note: Average quantities are the simple (unweighted) averages across rows. Corresponding error intervals are the union of those from individual studies.
* This estimate of *Average land area* includes all 12 rows above, it mixes explicit measurements of built area with others.
† This estimate was obtained by the author through visual inspection of Fig. 1 in Ref. (*39*).

Table S3: Summary of empirical evidence for predicted urban scaling exponents. Note that *Land area* has been measured in a variety of ways; those that account only for built area (at different levels of resolution) and those that compute circumscribing area. The latter scale with a smaller exponent $\alpha \sim 2/3$, while the former should approach the exponent $\nu \sim 5/6$, rather than α, as their resolution improves, as observed. Additional, measurements, discussion and sources are given in the original references and in the Supplementary Text.

References and Notes

1. L. Mumford, *The City in History: Its Origins, Its Transformations, and Its Prospects*. (Harcourt, Brace & World, Inc., New York, 1961).

2. I. Rodríguez-Iturbe, A. Rinaldo, *Fractal River Basins: Chance and Self-Organization*. (Cambridge Univ. Press, New York, 1997).

3. E. N. Bacon, *The Design of Cities* (Penguin Group USA, New York, 1976).

4. G. B. West, J. H. Brown, B. J. Enquist, A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122 (1997). doi:10.1126/science.276.5309.122 Medline

5. G. B. West, J. H. Brown, B. J. Enquist, The fourth dimension of life: fractal geometry and allometric scaling of organisms. *Science* **284**, 1677 (1999). doi:10.1126/science.284.5420.1677 Medline

6. H. Samaniego, M. E. Moses, Cities as organisms: Allometric scaling of urban road networks. *J. Transp. Land Use* **1**, 21 (Summer 2008).

7. J. S. Waters, C. T. Holbrook, J. H. Fewell, J. F. Harrison, Allometric scaling of metabolism, growth, and activity in whole colonies of the seed-harvester ant Pogonomyrmex californicus. *Am. Nat.* **176**, 501 (2010). doi:10.1086/656266 Medline

8. D. S. Dendrinos, H. Mullally, *Urban Evolution: Studies in the Mathematical Ecology of Cities.* (Oxford Univ. Press, Oxford, 1985).

9. L. Sveikauskas, The Productivity of Cities. *Q. J. Econ.* **89**, 393 (1975). doi:10.2307/1885259

10. E. L. Glaeser, B. Sacerdote, Why is there more crime in cities? *J. Polit. Econ.* **107**, S225 (1999). doi:10.1086/250109

11. L. M. A. Bettencourt, J. Lobo, D. Strumsky, Invention in the City: Increasing returns to patenting as a scaling function of metropolitan size. *Res. Policy* **36**, 107 (2007). doi:10.1016/j.respol.2006.09.026

12. L. M. A. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, G. B. West, Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7301 (2007). doi:10.1073/pnas.0610172104 Medline

13. M. Batty, The size, scale, and shape of cities. *Science* **319**, 769 (2008). doi:10.1126/science.1151419 Medline

14. See supplementary materials for details.

15. R. Anderson, R. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Oxford, 1991).

16. S. Angel, J. Parent, D. L. Civco, A. Blei, D. Potere, The dimensions of global urban expansion: Estimates and projections for all countries, 2000–2050. *Prog. Plann.* **75**, 53 (2011). doi:10.1016/j.progress.2011.04.001

17. D. Saunders, *Arrival City* (Pantheon Books, New York, 2010).

18. S. Nordbeck, Urban allometric growth. *Geogr. Ann.* **53**, 54 (1971). doi:10.2307/490887

19. M. Barthélemy, Spatial Networks. *Phys. Rep.* **499**, 1 (2011). doi:10.1016/j.physrep.2010.11.002

20. S. Milgram, The experience of living in cities. *Science* **167**, 1461 (1970). doi:10.1126/science.167.3924.1461 Medline

21. M. Schläpfer *et al*., *The Scaling of Human Interactions with City Size*; http://arxiv.org/abs/1210.5215.

22. J. Jacobs, *The Economy of Cities* (Vintage Books, New York, 1969).

23. A. Marshall, *Principles of Economics* (MacMillan, London, 1920). See book IV, chapter X.

24. E. L. Glaeser, J. D. Gottlieb, The wealth of cities: Agglomeration economies and spatial equilibrium in the United States. *J. Econ. Lit.* **47**, 983 (2009). doi:10.1257/jel.47.4.983

25. A. Gomez-Lievano, H. Youn, L. M. A. Bettencourt, The statistics of urban scaling and their connection to Zipf's law. *PLoS ONE* **7**, e40393 (2012). doi:10.1371/journal.pone.0040393 Medline

26. C. Alexander, A City is Not a Tree. *Archit. Forum* **122**, 58 (1965).

27. A. Downs, The law of peak-hour expressway congestion. *Traffic Quarterly* **16**, 393 409 (1962).

28. G. Duranton, M. A. Turner, The fundamental law of road congestion: Evidence from US Cities. *Am. Econ. Rev.* **101**, 2616 (2011). doi:10.1257/aer.101.6.2616

29. S. Sassen, *The Global City: New York, London, Tokyo* (Princeton Univ. Press, Princeton NJ, 1991).

30. M. Fujita, P. Krugman, A. J. Venables, *The Spatial Economy: Cities, Regions, and International Trade* (MIT Press, Cambridge MA, 2001).

31. F. Lederbogen *et al*., City living and urban upbringing affect neural social stress processing in humans. *Nature* **474**, 498 (2011). doi:10.1038/nature10190 Medline

32. J. Leskovec, J. Kleinberg, C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations." In *KDD '05 Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 177–187 (2005); http://dl.acm.org/citation.cfm?id=1081893.

33. The smallest conceivable city has population size $N = 2$, as I assume that a city is intrinsically a social network and, as such, is predicated on the existence of social interactions.

34. L. M. A. Bettencourt, H. Samaniego, H. Youn, Professional diversity and the productivity of cities; http://arxiv.org/abs/1210.7335.

35. S. Angel, J. Parent, D. L. Civco, A. M. Blei, "The persistent decline in urban densities: Global and historical evidence of sprawl." *Lincoln Institute of Land*

Policy Working Paper; www.lincolninst.edu/pubs/1834 The-Persistent-Decline-in-Urban-Densities.

36. W. Alonso, *Location and Land Use* (Harvard Univ. Press, Cambridge MA, 1964).

37. M. Fujita, *Urban Economic Theory* (Cambridge Univ. Press, Cambridge, 1989).

38. J. R. Banavar, A. Maritan, A. Rinaldo, Size and form in efficient transportation networks. *Nature* **399**, 130 (1999). doi:10.1038/20144 Medline

39. J. Q. Stewart, W. Warntz, Physics of Population Distribution. *J. Reg. Sci.* **1**, 99 (1958). doi:10.1111/j.1467-9787.1958.tb01366.x

40. W. R. Tobler, Satellite confirmation of settlement size coefficients. *Area* **1**, 30 (1969).

41. P. A. Longley, M. Batty, J. Shepherd, The size, shape and dimension of urban settlements. *Trans. Inst. Br. Geogr.* **16**, 75 (1991). doi:10.2307/622907

42. K. Paulsen, Yet even more evidence on the spatial size of cities: Urban spatial expansion in the US, 19802000. *Reg. Sci. Urban Econ.* **42**, 561 (2012). doi:10.1016/j.regsciurbeco.2012.02.002

43. L. M. A. Bettencourt, J. Lobo, D. Strumsky, G. B. West, Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. *PLoS ONE* **5**, e13541 (2010). doi:10.1371/journal.pone.0013541 Medline

44. M. J. Woldenberg, An allometric analysis of urban land use in the United States. *Ekistics* **36**, 282 (1973).

45. G. H. Dutton, Criteria of growth in urban systems. *Ekistics* **215**, 298 (1973).

46. W. J. Coffey, Allometric growth in urban and regional social-economic systems. *Can. J. Reg. Sci.* **11**, 49 (1979).

47. P. Sutton, D. Roberts, C. Elvidge, H. Meij, A comparison of nightime satellite imagery and population density in the United States. *Photogramm. Eng. Remote Sensing* **63**, 1303 (1997).

48. M. Batty, P. Ferguson, Defining city size. *Environment and Planning B* **38**, 753 (2011). doi:10.1068/b3805ed

49. M. See, Batty, P. A. Longley, *Fractal Cities: A Geometry of Form and Function* (Academic Press, London UK, 1994).

50. M. Guerois, *Les Formes des Villes Européenes vues du Ciel*, thesis, Universite Paris I Pantheon-Sorbonne (2003).

51. See figures 5 and 6 and associated analysis in (*63*)

52. Aristotle, *Politics, Book I*, 1-2.

53. A. Pred, *Urban Growth and City-Systems in the United States, 1940-1860* (Harvard Univ. Press, Cambridge MA, 1980).

54. A. Lösch, *Die Räumliche Ordnung Der Wirtschaft* (Gustav Fischer Verlag, Jena, Germany, 1940). Translated from German by W. H. Woglom, *The Economics of Location* (Yale Univ. Press, New Haven, CT, 1954).

55. W. Christaller, *Die Zentralen Orte in S¨uddeutschland* (Gustav Fischer Verlag, Jena, Gremany, 1933). Translated from German by C. W. Baskin, *The Central Places of Southern Germany* (Prentice-Hall, Englewood Cliffs, NJ, 1966).

56. P. Hall, *Cities in Civilization* (Pantheon Books, New York, 1998).

57. J. D. McCarthy, O. R. Galle, W. Zimmern, Population density, social structure, and interpersonal violence: An intermetropolitan test of competing models. *Am. Behav. Sci.* **18**, 771 (1975). doi:10.1177/000276427501800604

58. S. Szreter, G. Mooney, Urbanization, mortality and the standard of living debate: New estimates of the expectation of life at birth in nineteenth-century British cities. *Econ. Hist. Rev.* **51**, 84 (1998). doi:10.1111/1468-0289.00084

59. R. Florida, *The Rise of the Creative Class: And How Its Transforming Work, Leisure, Community and Everyday Life* (Perseus Books, New York, 2002).

60. C. S. Fischer, *To Dwell Among Friends: Personal Networks in Town and City* (Univ. of Chicago Press, Chicago, 1982).

61. B. Wellman, *Networks in the Global Village: Life in Contemporary Communities* (Westview, Boulder CO, 1999).

62. L. M. A. Bettencourt, Research Notes. Available upon request from the author.

63. Y. Chen, Characterizing growth and form of fractal cities with allometric scaling exponents. *Discrete Dyn. Nat. Soc.* **2010**, 194715 (2010). doi:10.1155/2010/194715