The Origins of Scaling in Cities

Luís M. A. Bettencourt

SFI WORKING PAPER: 2012-09-014

SANTA FE INSTITUTE

The origins of scaling in cities

Luís M. A. Bettencourt[1*]

[1] Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe NM 87501, USA.

*E-mail: bettencourt@santafe.edu

July 1, 2012

Cities are perhaps the ultimate expression of human sociality displaying at once humanity's greatest achievements and some of its most difficult challenges. Despite the increasing importance of cities in human societies our ability to understand them scientifically, and manage them in practice, has remained unsatisfactorily 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, which exist in similar but changing forms over a huge range of scales. Here, I show how cities may evolve following a small set of basic principles that operate locally and can explain how cities change gradually from the bottom-up. As a result I obtain a theoretical framework that derives the general open-ended properties of cities through the optimization of a set of local conditions. This framework is used to predict, in a unified and quantitative way, the average social, spatial and infrastructural properties of cities as a set of scaling relations that apply to all urban systems, many of which have been observed in nations around the world. Finally, I compare and contrast the structure and dynamics of cities to those of other complex systems that share some analogous properties.

Keywords: urban dynamics, scaling, social networks, infrastructure, complex systems

Cities have been an endless source of fascination throughout human history [1, 2]. One of their most extraordinary properties is that cities exist, in recognizable but changing form, over an enormous range of scales from small towns with a few people to the largest metropolitan areas, presently headed by Tokyo with over 35 million inhabitants. In the last decades increasing urbanization, involving most of the world's nations and billions of people, has brought the problem of understanding cities to the fore of both policy and research [3]. There has been a long tradition of seeking insight into the nature of cities through analogies to other physical and biological systems. Though compelling, such metaphors, drawn from sources as diverse as river networks [4], biological organisms [5, 6, 7, 8], insect colonies [1, 2, 9] or ecosystems [10], have remained limited in helping us understand and plan cities successfully.

Recently, analyses of new and more extensive data from many urban systems worldwide have begun to establish a series of general statistical regularities of cities as systematic nonlinear variations of urban quantities, Y, with city size, N, measured as population or land area[11, 12, 13, 14, 15, 8, 16, 17], often as scale invariant relations $Y(N) = Y_0 N^{\beta}$, where Y_0 and β are constants in N. These empirical scaling results suggest that, despite their apparent complexity, cities may actually be quite simple as their average properties may be set by just a few key parameters [13, 15]. However, a fundamental derivation of these scaling relations has been lacking. Here, I develop a general theoretical framework of the interplay between social and urban infrastructural networks imbedded in space and time. From this perspective, I show how cities emerge as co-located, scale-invariant social networks made possible by co-evolved infrastructure networks subject to general efficiency constraints.

The most important properties of cities arise from two effects: i) the concentration of people in space and time; and ii) more intense use of urban material infrastructure. Together, i) and ii) promote social contact and coordination, increasing the production rates of social quantities, such as wealth, innovation, crime, etc [13] (superlinear scaling, $\beta \simeq 1.15$) and allow for savings

in roads, cables, etc per capita as cities grow (sublinear scaling, $\beta \simeq 0.85$), see Fig. 1.

To show how these properties are the result of the same essential dynamics consider the simplest model of a city with land area A and population N. I write the interactions between people i, j in terms of a social network A_{ij}^k and assume that social interactions are local, over an area a_0 and have strength g_k (k parameterizes different types of social links). The locality of interactions changes the simplest result that the number of links in a network with N nodes scales as $\sim N^2$ (Metcalfe's law), leading instead to (see SI for details) $Y = G\frac{N^2}{A}$, with $G \equiv \bar{g} a_0 \ell$, where $\bar{g}$ is the average link strength and ℓ is the typical length travelled by people, information, etc. Each urban output, Y, has physical units set by g_k, but it is often 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. This concept is familiar from population biology [18] and is the basis of definitions of functional cities as metropolitan areas [19]. Population mixing translates into the cost of realizing interactions proportional to the transverse dimension of the city $L = A^{1/2}$. Thus, the power spent in transport processes to keep the city mixed is $W = \epsilon L N = \epsilon A^{1/2} N$, where ϵ is a force per unit time. This cost must be covered by each individual's budget, $y = Y/N$, requiring $y \simeq W/N$, which implies $A(N) = aN^\alpha$ with $\alpha = 2/3$ and $a = (G/\epsilon)^\alpha$. Thus, $Y = Y_0 N^\beta$, where $\beta = 2 - \alpha = 1 + \frac{1}{3} > 1$ and $Y_0 = G^{1-\alpha}\epsilon^\alpha$. This simple model satisfies both principles i) and ii) and leads to area A that scales *sublinearly* with N ($\alpha = 2/3 < 1$), and socioeconomic outputs, Y, scaling *superlinearly* ($\beta = 4/3 > 1$). However, in practice this gives only an upper bound on β. As cities grow, transportation of people, goods and information becomes channeled into networks, which reduce dissipation relative to direct unstructured paths and obey a distinct set of principles.

Although they take many diverse forms, the volume of all urban infrastructure networks tends to scale sublinearly with population size N [13, 8], but faster than total land area. We can arrive at these facts and obtain scaling laws consistent with data by requiring that cities

grow following four principles: *1. Mixing Population:* The city develops so that citizens can explore it fully, given the resources at their disposal. I formalize this principle relative to the sketch given above 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. Thus, this can be seen also as an entry condition into the city [20]. I characterize the geometry of paths through a Hausdorff dimension, H, so that distance travelled $\propto A^{\frac{2}{2+H}}$ (see SI). Matching density to cost I obtain a generalized area scaling relation, $A(N) = aN^{\alpha}$, $\alpha = \frac{2}{2+H}$ ($\alpha = \frac{D}{D+H}$ in D dimensions). $H \sim 1$ is special because it allows each individual to fully explore the city within the smallest distance travelled (see SI for discussion). *2. Incremental network growth:* Infrastructure networks develop to connect people as they join (leading to *decentralized* networks [8]). The average distance between individuals is $d = \rho^{-1/2} = \left(\frac{A}{N}\right)^{1/2}$. This implies the total network area $A_n(N) \sim s_* N d = s_* A^{1/2} N^{1/2}$, where s_* is an invariant $D-1$ dimensional volume characterizing the smallest network transverse dimension. Together with 1. this implies $A_n \sim s_* a^{1/2} N^{5/6}$. ($A_n(N) \sim s_* A^{1/D} N^{(D-1)/D} = s_* a^{1/D} N^{\frac{D^2+D-1}{D(D+1)}}$ in D dimensions). *3. Bounded human effort:* The coupling $G = a_0 \bar{g} \ell$ is an approximate constant of city size. This includes (see SI) the interaction area per individual a_0, the quality and variety of social interactions $\bar{g}$, and the travel distance ℓ per person. As I show below, microscopic models of the city determine an optimal $G = G^*$, so that certain aspects of this assumption follow from others. Finally, *4. Socioeconomic outputs are proportional to the number of locally interacting people*, so that $Y = GN^2/A_n \sim N^{1+1/6}$, which yields scaling exponents in agreement with a wide variety of data [11, 12, 13, 14, 17, 21, 22, 23].

These principles do not require an explicit realization of networks, only their average properties. However, we can gain further insight by building a more microscopic theory of hierarchical organization of infrastructure networks (c.f. [4, 6, 7]). This requires stronger assumptions but will also reveal how some of the principles above follow from the general structure of interac-

tions and dissipation. Consider a hierarchical network with branching b and number of levels h. b measures the average ratio of infrastructure at successive levels, e.g. paths to small roads, larger roads to highways. Note that the structure of these networks is not a simple hierarchical tree [24]. The length of a network segment at level i is l_i and its transverse dimension s_i.(an area in 3D networks and a length in 2D, e.g. roads). Because 2. requires that the total network length is area filling $l_i = a_i/l$, with $a_i = ab^{i(\alpha-1)}$. The total network length L_n and area A_n are

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

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

where $N_i = b^i$, $N_h = N$ and $\delta = \frac{H}{D^2}\alpha$. I assumed $\alpha + \delta < 1$, which holds for $D > 1$. To obtain this result I took the transverse dimension of network terminal units, s_*, to be an invariant. This imposes a boundary condition on $s_i = s_* b^{(i-h)\sigma}$, where $\sigma = \delta - 1 < 0$ and thus $s_0 = s_* b^{-h\sigma} >> s_h = s_*$, so that the width of network sections at larger scales is much larger than for small ones, e.g. the width of highways versus small paths. Contrary to biological vascular systems, where both the length and radius of the network at the smallest scales was assumed invariant [7, 6], there is no evidence in cities for the invariance of the former, which depends on the detailed layout of the city (e.g. longer or shorter blocks, larger or smaller buildings, etc). These relations also imply that the network average transverse dimension $\bar{S} = A_n/L_n \sim N^{1/6}$.

The scaling of transverse dimensions together with the average conservation of flux in the network $s_i \rho_i v_i N_i = s_{i-1} \rho_{i-1} v_{i-1} N_{i-1}$ for all i, sets the scaling for $\rho_i v_i$, which is the average current cross-sectional density in each network branch. This is interesting because it characterizes the speed and density of carriers in the network at different levels, which controls dissipation mechanisms in the city. As a consequence of flux conservation and of the scaling of s_i, I obtain that $\rho_i v_i = b^{-\delta} \rho_{i-1} v_{i-1}$, which implies that the current density decreases from the root to the leaves, so that e.g. highways are faster and more densely packed than smaller roads, as observed

[25]. Making the further assumption that the current density at the smallest branches is invariant $\rho_h v_h = \rho_* v_*$ this leads to $\rho_i v_i = b^{\delta(h-i)}\rho_* v_*$. Then, because the total current is conserved it is independent of the level i and takes the value $I_i = I = I_0 N$, with $I_0 = s_*\rho_* v_*$. I now derive the properties of dissipative processes in the network. There are many possible forms of dissipation, including those that occur at large velocity or density. I make the standard assumption that the resistance per unit length per transverse network area, r, is constant [4, 7], leading to $r_i = r\frac{l_i}{s_i}$. For N_i parallel resistors this gives the total resistance per level $R_i = \frac{ar}{ls_*}b^{-i(1-\alpha+\delta)+h\sigma}$ and the total dissipated power, W, is

$$W = \sum_{i=1}^{h} W_i = I^2 \sum_{i=1}^{h} R_i = rI^2 \frac{a}{ls_*} b^{h\sigma} \frac{1 - b^{-(h+1)(1-\alpha+\delta)}}{1 - b^{-(1-\alpha+\delta)}} \simeq W_0 N^{1+\delta}, \tag{3}$$

which scales *superlinearly* with N, with $1 + \delta \simeq 1 + 1/6$ and $W_0 \simeq r\frac{as_*(\rho_* v_*)^2}{l(1-b^{-(1-\alpha+\delta)})}$. Thus, dissipation scales naturally like social interactions revealing the fundamental nature of cities as scale-invariant complex adaptive systems. Finally, we cast the problem of defining cities in terms of standard optimization, maximizing social outputs subject to network dissipation, as

$$\mathcal{L} = Y - W + \lambda_1 \left(\epsilon A^{H/D} - GN/A\right) + \lambda_2 \left(A_n - c\rho^{-1/D}N\right) \rightarrow \frac{2\alpha - 1}{\alpha} G^* \frac{N^2}{A_n(N)}, \tag{4}$$

where c is a constant, see SI text, and λ_1, λ_2 are Lagrange multipliers. Eq. 4 gives the basis for the derivation of the properties of every segment in the network, through Eqs. 1-2. This results in the scaling of area, network properties and socioeconomic quantities derived above, see Table 1 for a summary. The novelty in Eq. (4) is the prediction of an optimal G, through $d\mathcal{L}/dG = 0$, see Fig.2b. Both socieconomic outputs and dissipation grow with G but the latter grows with a larger power, leading to two solutions to $\mathcal{L} = 0$: $G = G_{\min} = 0$ and $G = G_{\max} = \left[\frac{(\epsilon^\alpha l)^2}{r' I_0^2}\right]^{\frac{1}{2\alpha-1}}$, where $r' \simeq r$, see SI. Thus, if the balance of social interactions is positive, $\bar{g} > 0$, human societies are always unstable towards the formation of cities. However, there is an upper $G = G_{\max}$ (reached e.g. via increases in human capital or mobility) beyond

which dissipation overcomes social benefits and the city becomes unstable. In between there is an optimum $G = G^* = \left[\frac{1-\alpha}{\alpha}\right]^{\frac{1}{2\alpha-1}} G_{\max} \leq G_{\max}$ where the city is most productive. The existence of G^* expresses the average balance between social interactions and infrastructure networks that defines the city at all scales.

The present conceptualization of cities deals primarily with average quantities (mean field theory). An eventual formalization that includes statistical fluctuations [21, 22, 23] and details of specific quantities, Y, will contribute to a more complete urban theory and may improve on the prediction of the value of particular exponents. There are several interesting analogies between this view of cities and other complex systems. As in biological organisms [6, 7], infrastructure networks are volume filling; however in cities they scale faster with population than their embedding volume. Thus, infrastructure in large cities often moves into the third dimension, above or below ground. In contrast, by concentrating people and their interactions as sources of the system's productivity, the city leads to increasing returns to population ($\beta > 1$), magnifying per capita wealth creation, innovation and crime and accelerating all forms of social life [13]. This effective contraction of time is often observable in the acceleration of particular behaviors [13] and may be associated with increased cognitive stimulation and stress [26, 27]. As such, cities manifest the *opposite* character to biological organisms, where the target of optimization is the minimization of energy dissipation [4, 6, 7] and the pace of change is determined by network constraints making larger organisms slower. This makes transport in larger urban networks less efficient, which is necessary to enable the growth in the city's primary social functions [28]. It is because dissipative processes scale like social interactions that cities can be scale invariant as larger cities can incur the same average costs per unit of productivity as small towns, while growing in functional diversity [28].

This acceleration and concentration of interactions in cities has parallels in other systems that are driven by attractive forces and that become denser with scale. The simplest such sys-

tems are stars, whose luminosity (power emitted) increases superlinearly with mass. However, differences in the nature of interactions and transport processes make their scaling exponents different quantitatively and ultimately limit the complexity that a star can achieve. It is a fascinating question if networks that densify with increasing scale [29], from ecosystems to information networks in biology and society, share any similarities with cities despite their different relationships to physical space.

In summary, I showed how the general scaling properties of cities can be derived from a set of local principles that account for their gradual development, from the bottom-up, as densifying social networks subject to geometric and efficiency constraints on urban infrastructure networks. This theoretical framework shows how the many social, infrastructural, spatial and temporal aspects of the city are entangled and expresses their common origin in terms of fundamental simpler dynamics. Unveiling these deeper connections is a necessary step towards a more scientific approach to urban planing and policy and may shed light on some of the essential conditions that have lead to the extraordinary development of complex human societies.

References

[1] Aristotle, *Politics, Book I*, 1-2.

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

[3] UN-HABITAT (United Nations Human Settlements Program), *State of the Worlds Cities 2010/11*. Available online at `http://www.unhabitat.org`

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

[5] E. N. Bacon, *The Design of Cities* (Penguin Group USA, New York, NY, 1976).

[6] G. B. West, J. H. Brown, B. J. Enquist, A General Model for the Origin of Allometric Scaling Laws in Biology, *Science* **276**, 122-126 (1997).

[7] G. B. West, J. H. Brown, B. J. Enquist, The Fourth Dimension of Life: Fractal Geometry and Allometric Scaling of Organisms. *Science* **284**, 1677-1679 (1999).

[8] H. Samaniego, M. E. Moses, Cities as organisms: Allometric scaling of urban road networks. *J. Transp. Land Use*, **1**, 2139 (2009).

[9] J. S. Waters, C. T. Holbrook, J. H. Fewell and J. F. Harrison, Allometric Scaling of Metabolism, Growth, and Activity in Whole Colonies of the Seed-Harvester Ant Pogonomyrmex californicus. *The American Naturalist* **176**, 501-510 (2010); A. I. Bruce, M. Burd, Allometric scaling of foraging rate with trail dimensions in leaf-cutting ants. *Proc. R. Soc. B* **279**, 2442-2447 (2012).

[10] D. S. Dendrinos, H. Mullally, *Urban Evolution: Studies in the Mathematical Ecology of Cities* (Oxford University Press, Oxford, 1985).

[11] L. Sveikauskas, The Productivity of Cities. *Q. J. Econ.* **89**, 393-413 (1975).

[12] E. L. Glaeser, B. Sacerdote, Why is There More Crime in Cities?. *J. of Polit. Econ.* **107**, S225-S258 (1999).

[13] 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**, 7301-7306 (2007).

[14] L. M. A. Bettencourt, J. Lobo, D. Strumsky, Invention in the City: Increasing Returns to Patenting as a Scaling Function of Metropolitan Size. *Res. Pol.* **36**, 107-120 (2007).

[15] M. Batty, The Size, Scale, and Shape of Cities. *Science* **319**, 769 -771 (2008).

[16] M. A. Changizi, M. Destefano, Common Scaling Laws for City Highway Systems and the Mammalian Neocortex. *Complexity* **15**, 11-18 (2009).

[17] E. L. Glaeser, J. D. Gottlieb, The Wealth of Cities: Agglomeration Economies and Spatial Equilibrium in the United States. *Journal of Economic Literature*, **47**, 983-1028 (2009).

[18] R. Anderson, R. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, Oxford, 1991).

[19] See e.g. `http://www.census.gov/population/metro/`

[20] D. Saunders, *Arrival City* (Pantheon Books, New York NY, 2010).

[21] 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).

[22] A. Gomez-Liévano, H. Youn, L.M.A Bettencourt, The Statistics of Urban Scaling and their Connection to Zipf's Law. Santa Fe Institute working paper 12-02-001. Available online at `http://www.santafe.edu/media/workingpapers/12-02-001.pdf`. To appear in *PLoS One*.

[23] M. Schläpfer, *et al.* The Scaling of Human Interactions with City Size. *in review* (2012).

[24] C. Alexander, A city is not a tree, *Architectural Forum*, **122**, 58-62 (1965).

[25] A. Downs, The Law of Peak-hour Express-way Congestion. *Traffic Quarterly* **16**, 393 409 (1962); G. Duranton, M. A. Turner, The Fundamental Law of Road Congestion: Evidence from US Cities. *American Economic Review* **101**, 2616-2652 (2011).

[26] S. Milgram, The Experience of Living in Cities, *Science* **13**, 1461-1468 (1970).

[27] F. Lederbogen *et al.*, City Living and Urban Upbringing Affect Neural Social Stress Processing in Humans, *Nature* **474**, 498-501 (2011).

[28] J. Jacobs, *The Death and Life of Great American Cities* (Random House, New York NY, 1961).

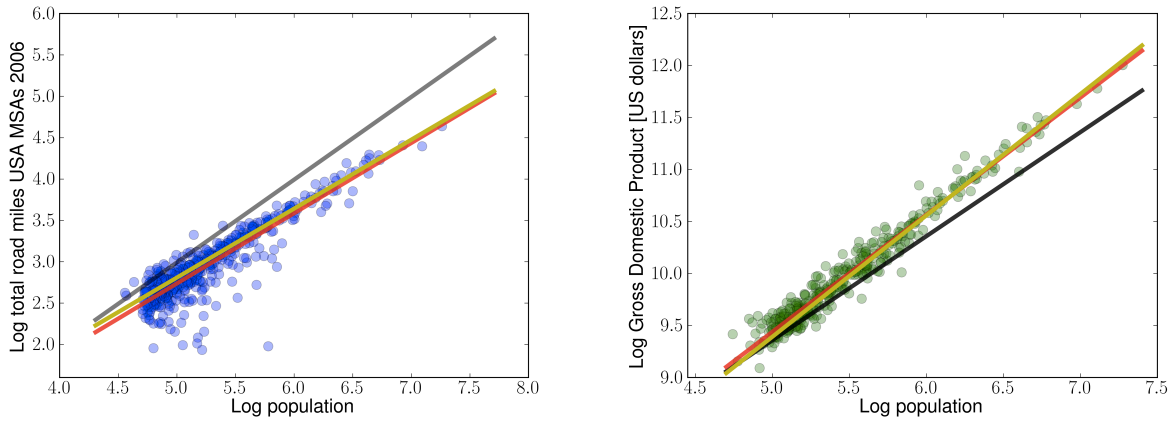
[29] J. Leskovec, J. Kleinberg, C. Faloutsos, Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. in *KDD '05 Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, p 177 - 187 (2005). Available online at `http://dl.acm.org/citation.cfm?id=1081893`

Urban Scaling Relations	Model (D=2,H=1)	Model (D, H)	Effect
Land area $A = aN^{\alpha}$	$\alpha = \frac{2}{3}$	$\alpha = \frac{D}{D+H}$	spatial densification
Network volume $A_n = A_{n0}N^{\nu}$	$\nu = \frac{5}{6}$	$\nu = 1 - \delta = \frac{D^2+DH-H}{D(D+H)}$	growth of infrastructure
Network length $L = L_0 N^{\lambda}$	$\lambda = \frac{2}{3}$	$\lambda = \alpha$	area filling networks
Average network width $\bar{S} = \bar{S}_0 N^{\bar{\sigma}}$	$\bar{\sigma} = \frac{5}{6}$	$\bar{\sigma} = 1 - \delta$	widening of roads
Interactions per capita $y = Y_0 N^{\delta}$	$\delta = \frac{1}{6}$	$\delta = \frac{H}{D(D+H)}$	increased interactions
Socioeconomic rates $Y = Y_0 N^{\beta}$	$\beta = \frac{7}{6}$	$\beta = 1 + \delta = \frac{D^2+DH+H}{D(D+H)}$	acceleration of social rates
Power dissipation $W = W_0 N^{\omega}$	$\omega = \frac{7}{6}$	$\omega = 1 + \delta$	increased congestion
Land Value $P_L = P_0 N^{\delta_L}$	$\delta_L = \frac{3}{2}$	$\delta_L = 1 + \alpha + \delta$	increased land rents

Table 1: Urban indicators and their scaling relations. The first column shows expected mean-field values for scaling exponents vs. population size ($D = 2, H = 1$). The second column shows the value of scaling quantities in general D spatial dimensions. The third column describes the effect.

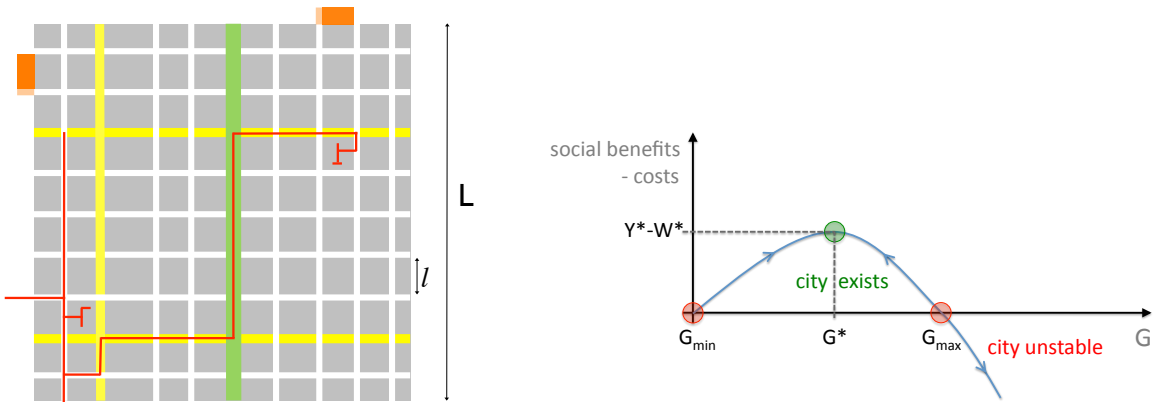
Acknowledgements

I thank José Lobo, Geoffrey West and HyeJin Youn for discussions. This research is partially supported by the Rockefeller Foundation, the James S. McDonnell Foundation (grant no. 220020195), the National Science Foundation (grant no. 103522), the John Templeton Foundation (grant no. 15705) and by a gift from the Bryan J. and June B. Zwan Foundation.



(a) (b)

Figure 1: Scaling of urban infrastructural and socioeconomic quantities. (a) Total lane miles (volume) of roads in US metropolitan areas in 2006 (blue dots). Lines show the best fit to a scaling relation $Y = Y_0 N^{\beta}$ (red), with $\beta = 0.849 \pm 0.038$ (95% CI, $R^2 = 0.65$), the theoretical prediction for $\beta = 5/6$ (yellow) and linear scaling $\beta = 1$ (black). (b) Gross Metropolitan Product of US metropolitan area in 2006 (green dots). Lines show the best fit (red), with $\beta = 1.126 \pm 0.023$ (95% confidence interval, $R^2 = 0.96$), the theoretical prediction, $\beta = 7/6$ (yellow), and proportional scaling, $\beta = 1$ (black).



(a) (b)

Figure 2: The city in space and its balance of social and dissipative processes. (a) Grey blocks denote settled areas while spaces in between (white, yellow, green) represent infrastructure networks. Note that land area $A = L^2$, and that network length L_n is proportional to A, $L_n = 2(\lambda + 1)L \simeq A/l$, see SI for details. Red lines denote the volume of public space spanned by an individual, which determines his mean number of social interactions and productivity. As the city grows and new land is settled (orange blocks) the infrastructure network grows incrementally (pale orange). (b) There is an optimal value of the parameter G at which city's are most productive. Social networks become unstable to the formation of cities if net social interactions are positive $G > G_{\min} = 0$, and less than an upper value $G < G_{\max}$ (red circles), when costs overcome social benefits. In between there is an optimal $G = G^*$ (green circle) at which benefits Y^* maximally overcome costs W^*.

Supplementary Information
The origins of scaling in cities

Luís M. A. Bettencourt[1*]

[1] Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe NM 87501, USA.

[*]E-mail: bettencourt@santafe.edu

July 1, 2012

Supplementary Text

From social interaction networks to the mean field approximation

In general I can write the sum total of all social interactions in terms of a generalized graph, A_{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

$$Y = \sum_{i,j;k} g_k A_{ij}^k, \tag{S.1}$$

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,

$$I_{i,k} = \sum_j A^k_{ij}. \tag{S.2}$$

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 $A^k_{ij} = p(k|ij)A_{ij} = p(k)A_{ij}$, where $p(k)$ is the probability of different interaction modes, k, per link and A_{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), which I assume is an invariant (a property of a person, independent of population size or other urban quantities) and by a length traveled in the network ℓ. This spans a *world-sheet*, which is a fraction of the total public space volume, V_n, (or area, in 2D networks) 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 total interactions experienced by our test individual are given by the ratio of the two

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

$$\begin{aligned}\bar{I}_{i,k} &= p(k)\int d^D x \rho(x)\delta\left(x - x(t)\right) \simeq p(k)\rho_n \int dx^D \delta\left(x - x(t)\right) \\ &= p(k)a_0\ell\frac{N-1}{V_n} \simeq p(k)a_0\ell\frac{N}{V_n},\end{aligned} \tag{S.3}$$

with ρ_n the average density in public networks, where interactions take place. In this last expression I wrote $N - 1 \simeq N$, for large N (*1*). Note that any other sufficiently short-range potential, not necessarily a δ-function, would lead to the same result, Eq. (S.3), up to a dimensionless multiplicative constant, independent of N. Then I can finally write

$$\bar{I}_i = \sum_k \bar{I}_{i,k} = \bar{g}\frac{a_0\ell}{V_n}N, \qquad \bar{g} = \sum_k p(k)g_k, \tag{S.4}$$

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

$$Y = \sum_{i=1}^{N} \bar{I}_i = G\frac{N^2}{A_n}, \tag{S.5}$$

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. (S.1) 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, while others may benefit from the mean-field effects that result from being in the wider city and collecting a diversity of interactions. This analysis of the finer structure of more heterogeneous interactions, which requires considerations beyond the average behavior derive 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 detail in (*2*).

Mixing, exploration of space and Hausdorff 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. (S.6) by H, the Hausdorff dimension of a path in space. C_m is the cost associated with such path, which is written in terms of the city's land area, A, as

$$C_m = \epsilon A^{\frac{H}{D}}, \tag{S.6}$$

in D general dimensions. 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 C_m to $Y_{\min}$ leads to a relationship between population

and area of the form

$$A = aN^{\alpha}, \tag{S.7}$$

with the exponent $\alpha = D/(D + H) \simeq 2/3$, for $D = 2$, $H = 1$, and the area per person $a = (G/\epsilon)^{\alpha}$. 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, leading to 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 throughout the world to become less dense (*3*). 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.

$H = 1$ corresponds to the most natural assumption, that these costs are proportional to the linear extent 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 (*4, 5*).

$H < 1$ 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 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 as 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. Given available data, which provides typically only total administrative unit area for a city or metropolitan area, this $H < 1$ regime seems to be sometimes observed, see

discussion below.

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 $\alpha \geq 1/2$. 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 here 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 manifestly counterfactual, certainly for large cities. For all these reasons, while I leave H as a parameter in the main tex, I expect that it would naturally be of order $H \simeq 1$, with $\alpha \simeq 2/3$ as is observed for contemporary US metropolitan areas.

There have been many attempts at characterizing the scaling relation between the land area of cities and their population. Most of these characterizations use definitions of cities in terms of sets of administrative units (counties, municipalities) which leads to several potential biases. Nevertheless, for example Nordbeck (*6*) found that for cities in Sweden over two time periods $\alpha \simeq 2/3$, which is also consistent with US metropolitan statistical areas in recent years (*7*). However, other studies found values of α in the range $2/3 \leq \alpha < 1$, (*8*), but these report on a variety of different definitions of city, ranges of scales, etc. While increases of population density with city size within each urban system and city type are an undisputed property of cities worldwide more consistent and accurate measurements of the scaling of land area with population remain necessary. Note, finally, that under the assumption of decentralized infrastructure networks, land area, A, enters the determination of A_n via a factor of $A^{1/D}$, so that the impact of fluctuations in A is reduced as they enter other urban quantities and differences in the value of α are effectively halved ($D = 2$).

Infrastructure networks' length is area filling

I have assumed in the main text the property that networks of infrastructure fill the occupied 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 1 illustrates this situation for a regular grid. In this case the total length of the network can be derived easily, see Figure 1 (main text), as

$$A = L^2 = (\lambda l)^2; \qquad L_n = 2(\lambda + 1)L = 2(\lambda + 1)\lambda l = \frac{2}{l}A + 2\sqrt{A} \underset{\lambda >> 1}{\sim} A, \tag{S.8}$$

where l is the average block size, λ is the number of blocks in the city, and $L = \lambda l$. 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 $L_n \sim A/l$.

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 invariant network properties. This choice is important because it sets the scaling behavior of dissipation. I have assumed that the width of terminal 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 that does not vary with city population size. This means that I can write the scaling of width across network levels as

$$s_i = s_* b^{(i-h)(\delta-1)}, \tag{S.9}$$

which implies that the width is largest at the highest level ($i = 0$: root, "highways") $s_0 = b^{h(1-\delta)}$, 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 flux that

$$s_i \rho_i v_i N_i = s_{i-1} \rho_{i-1} v_{i-1} N_{i-1}, \qquad \forall_{i=1}^{h}. \tag{S.10}$$

This condition should be relaxed in general for a network that is not a (balanced) tree, as for example, would happen in a semi-lattice (*9*), where branches at the same level are connected, or upper branches can converge on the same lower site. Note that this is the only place where the tree structure of urban networks is used explicitly. The difficulty with these generalized structures is that their geometry can be highly variable and the result must be treated statistically, or just bounded (*10*). The assumption made here, as we shall show below, leads to the smallest dissipation. Any of these generalized structures would therefore be more restrictive. This condition leads to the scaling relation for the current density

$$\rho_i v_i = b^{-\delta} \rho_{i-1} v_{i-1}. \tag{S.11}$$

This relationship is not fully specified until we prescribe its boundary conditions. We can place a limit on the current density at the root $\rho_0 v_0 = \text{const}$, which leads to $\rho_i v_i = b^{-i\delta} \rho_0 v_0$, 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

$$I_i = s_i \rho_i v_i N_i = s_* \rho_0 v_0 b^{h(1-\delta)}, \tag{S.12}$$

or

$$I_i = s_* \rho_* v_* b^h, \tag{S.13}$$

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 we have $I_i = I = s_* \rho_0 v_0 N^{1-\delta}$, and for the boundary condition at the leaves this leads to $I = I_0 N$, with $I_0 = s_* \rho_* v_*$. Note that the latter is the expected current for a population of identical individuals, in terms of their intrinsic parameters, and is therefore the natural boundary condition. It means in intuitive terms that the flow of people through doors 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 (*11*). Thus, the differences between cities arise at larger scales, where social interactions are more common and population-wide constraints apply, see below.

Dissipation on infrastructure networks

There are many dissipative processes (costs) that can take place in a city and that can lead to situations where increasing social interactions and their products may be more than overcome by their associated costs. In the main text we assume the the resistance at each level of the network is that of all branches taken in parallel (c.f (*12*)), that is

$$R_i = \left[\sum_{i=1}^{N_i} \frac{1}{r_i} \right]^{-1} = \frac{r_i}{N_i}, \tag{S.14}$$

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,

$$r_i = r \frac{l_i}{s_i} = r \frac{a}{l s_*} b^{(\alpha-\delta)i + h\sigma}, \tag{S.15}$$

which increases with level i, and is therefore larger at the smallest branches than at the root. From (S.14) this leads to

$$R_i = r \frac{a}{l s_*} b^{-(1-\alpha+\delta)i + h\sigma}, \tag{S.16}$$

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

$$W_i = R_i I^2, \tag{S.17}$$

$$W = \sum_{i=1}^{h} W_i = I^2 \sum_{i=1}^{h} R_i = r I^2 \frac{a}{l s_*} b^{h\sigma} \frac{1 - b^{-(h+1)(1-\alpha+\delta)}}{1 - b^{-(1-\alpha+\delta)}} = W_0 N^{1+\delta}, \tag{S.18}$$

which scales superlinearly, with an exponent $1 + \delta \simeq 7/6$. $W_0 \simeq r \frac{a s_* (\rho_* v_*)^2}{l(1-b^{-1+\alpha-\delta})}$. We see that the dissipative behavior of the network is set by the current squared multiplied by the resistance at the root. 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 bottlenecks that occur at its largest scales. Remarkably, this result ties together the most microscopic needs and behaviors of individuals to the most macroscopic aspects of the urban infrastructure.

Another way to see this is to rearrange terms in Eq. (S.18) to write it as

$$W = r' \left(\frac{a}{l}\right)^2 \frac{I^2}{A_n} \tag{S.19}$$

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 individual welfare. We write the objective function, $\mathcal{L}$, for this problem as

$$\mathcal{L} = Y - W + \lambda_1 \left(\epsilon A^{H/D} - GN/A\right) + \lambda_2 \left(A_n - c\rho^{-1/D}N\right). \tag{S.20}$$

where $c = s_* a^{1-1/D}/[l(1-b^{\alpha+\delta-1})]$ is a constant that follows from Eqs. 2 and 4 and λ_1, λ_2 are Lagrange multipliers. From the point of view of individuals, they 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 can provide organizations (which affect the interaction modes) and infrastructure so that their general socioeconomic benefits are maximized. This can be expressed in terms of the variation of A_n (and of the factors that make it). Varying (S.20) relative to A and A_n leads to

$$Y(N) = G\frac{N^2}{A_n(N)}, \qquad W(N) = r'\left(\frac{a}{l}\right)^2 \frac{I^2}{A_n(N)}, \tag{S.21}$$

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

$$G \equiv G_{\min} = 0, \qquad \text{or} \qquad G \equiv G_{\max} = \left[\frac{(\epsilon^\alpha l)^2}{r' I_0^2}\right]^{\frac{1}{2\alpha-1}}. \tag{S.22}$$

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 may loose population and even collapse. 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 $\mathcal{L}$ relative to G, (recall that $a = (G/\epsilon)^\alpha$), to obtain

$$\frac{d\mathcal{L}}{dG} = \left[(1-\alpha) - \alpha \frac{r' I_0^2}{G} \left(\frac{a}{l}\right)^2\right] \frac{N^2}{A_n(N)} = 0, \tag{S.23}$$

which results in the solution

$$G = G^* = \left[\frac{1-\alpha}{\alpha}\right]^{\frac{1}{2\alpha-1}} G_{\max} \leq G_{\max}. \tag{S.24}$$

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 dissipation. 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 the Lagrangian at G^* as

$$\mathcal{L} = Y - W = \frac{2\alpha - 1}{\alpha} G^* \frac{N^2}{A_n}. \tag{S.25}$$

We could in general consider a last step to optimize the problem over the exponent α (or conversely H, at fixed D) to find its optimal value. This optimization step is complex however and requires some further consideration of the microscopic aspects of the problem, so I leave it for future work. We can also turn (S.20) into a much more detailed optimization problem, by specifying Y, W, and A_n in terms of their detailed microscopic components, involving social and infrastructural networks, as discussed in the main text. This leads to variations relative to l_i, s_i, a_i, and $\rho_i v_i$. However the aggregate results on scaling remain unchanged. In principle

this procedure can be used in general to determine an optimal detailed structure of social networks, given infrastructure and vice-versa so that the balance of social outputs minus costs are maximized.

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 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. Note that taken separately, social output could be increased by e.g. increasing the magnitude of G, or reducing A_n. However, this would lead to larger increases in dissipation which would eventually overtake the original benefits. Conversely, as it has been often tried, dissipation can be minimized by infrastructure expansion and by other measures that decrease density. Such measures however tend to reduce social interactions thereby decreasing socioeconomic outputs. Thus, cities can exist in an open-ended state of dynamical equilibrium, characterized in principle by unlimited population and socioeconomic growth, provided an optimal interplay between social networks and infrastructure is maintained as they grow.

Data sources

Data for roads in United States Federal-Aid Urbanized Areas (mostly equivalent to Metropolitan Statistical Areas) is provided by the Office of Highway Policy Information from the Federal Highway Administration. It is available online at `http://www.fhwa.dot.gov/policy/ohim/hs05/roadway_extent.htm`. Gross Metropolitan Product for US Metropolitan Statistical Areas (functional cities) is compiled by the US Bureau of Economic Analysis (BEA) and is available online at `http://www.bea.gov/regional/`

References and Notes

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

2. L. M. A. Bettencourt and H. Samaniego, Professional diversity and the productivity of cities, *in review*.

3. 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*. Available online at `http://www.lincolninst.edu/pubs/1834_The-Persistent-Decline-in-Urban-Densities`

4. W. Alonso, *Location and Land Use* (Cambridge MA: Harvard University Press, 1964).

5. M. Fujita, *Urban Economic Theory* (Cambridge University Press: Cambridge, 1989).

6. S. Nordbeck, Urban allometric growth. *Geografiska Annaler* **53**B, 54-67 (1971).

7. J. Lobo and H. Youn, *private communication*.

8. See e.g. M. J. Woldenberg, An allometric analysis of urban land use in the United States. *Ekistics* **36**, 282-290 (1973); J. Q. Stewart, W. Warntz, Physics of Population Distribution, *Journal of Regional Science* **1**, 99-123 (1958); G. H. Dutton, Criteria of Growth in Urban Systems, *Ekistics* **215**, 298-306 (1973), W. J. Coffey, Allometric growth in urban and regional social-economic systems, *Canadian Journal of Regional Science* **11**, 49-65 (1979); M. Batty, P. A. Longley, Fractal Cities: A Geometry of Form and Function, (Academic Press, London, 1994).

9. C. Alexander, A city is not a tree, *Architectural Forum* **122**, 58-62 (1965).

10. J. Banavar, A. Maritan, A. Rinaldo, Size and Form of Efficient Transportation Networks. *Nature* **399**, 130-132 (1999).

11. 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**, 7301-7306 (2007).

12. G. B. West, J. H. Brown, B. J. Enquist, A general model for the Origin of Allometric Scaling Laws in Biology, *Science* **276**, 122-126 (1997).