nature human behaviour

LETTERS

PUBLISHED: 22 DECEMBER 2016 | VOLUME: 1 | ARTICLE NUMBER: 0012

Explaining the prevalence, scaling and variance of urban phenomena

Andres Gomez-Lievano[1]*, Oscar Patterson-Lomba[2] and Ricardo Hausmann[1,3,4]

The prevalence of many urban phenomena changes systematically with population size[1]. We propose a theory that unifies models of economic complexity[2,3] and cultural evolution[4] to derive urban scaling. The theory accounts for the difference in scaling exponents and average prevalence across phenomena, as well as the difference in the variance within phenomena across cities of similar size. The central ideas are that a number of necessary complementary factors must be simultaneously present for a phenomenon to occur, and that the diversity of factors is logarithmically related to population size. The model reveals that phenomena that require more factors will be less prevalent, scale more superlinearly and show larger variance across cities of similar size. The theory applies to data on education, employment, innovation, disease and crime, and it entails the ability to predict the prevalence of a phenomenon across cities, given information about the prevalence in a single city.

Scaling is ubiquitous across many phenomena[5], including physical[6] and biological[7] systems, plus a wide range of human[8,9] and urban activities[1,10]. Figure 1 shows, for US Metropolitan Statistical Areas, ten different phenomena classified into five broad types: employment, innovation, crime, educational attainment and infectious disease. We observe scaling in the sense that the counts of people engaged in (or suffering from) each phenomenon scale as a power of population size. This relation takes the form $E\{Y|N\} = Y_0 N^{\beta}$, where $E\{\cdot|N\}$ is the expectation operator conditional on population size N, Y is the random variable representing the 'output' of a phenomenon in a city, Y_0 is a measure of general prevalence of the activity in the country and β is the scaling exponent, that is, the relative rate of change of Y with respect to N. From Fig. 1, we can also observe notable differences in the average prevalence, the slopes of the regression lines and the variance across all ten phenomena. Hence, we seek to explain four empirical facts: prevalence follows a power-law scaling with population size, and different phenomena have different general prevalence, scaling exponents and variance for cities of similar size. Remarkably, these observations seem to be pervasive across phenomena, as we find them to be present in more than 40 different urban activities. Here we propose a mechanism to explain them simultaneously.

Scaling laws are important in science because they constrain the development of new theories: any theory that attempts to explain a phenomenon should be compatible with the empirical scaling relationships that the data exhibit. A number of mechanisms have been proposed to explain the origins of scaling. Most theories are based on a network description of the underlying phenomena and derive the scaling properties from the way in which the number of links grows with the number of nodes in the network, under some energy or budget constraints[11–16]. Other scaling relationships are the result of how lines relate to surfaces, and surfaces to volumes[17–20]. We propose a different mechanism that improves on previous explanations in that it not only generates scaling, but also accounts for the value of the scaling exponent, the average relative prevalence across different phenomena, and the variance within phenomena across cities of similar size.

The central assumption of our framework is that any phenomenon depends on a number of complementary factors that must come together for it to occur. More complex phenomena are those that require, on average, more complementary factors to be simultaneously present. This assumption is the conceptual basis for the theory of economic complexity[2,3,21,22].

In addition, as with models of cultural evolution, we posit that the number of factors in the environment is a function of population size[4,23,24]. Anthropological studies have shown this to be true about the diversity of skills, behaviours, beliefs, vocabulary and tools[25–30]. More recent evidence of this relationship has been found in cities[31–33]. These models assume that cultural accumulation is a Darwinian process, in the sense that it involves inheritance, differential fitness and selection. The prediction is a logarithmic function of population size[4]. Our approach is not dependent on the precise justification for the logarithmic function, since logarithms typically emerge from the fact that selection implies transforming initial distributions into extreme value distributions (such as a Gumbel distribution[4]) whose means grow logarithmically with sample size. For example, we can assume that each factor has a different probability of appearance, and that cities randomly sample from this distribution according to their size. If there is a process of selection, an extreme value distribution will emerge. In this setting, the diversity of factors will accumulate logarithmically with population size if the distribution of frequencies of the factors is Gumbel, meaning that the rarer factors will only appear in larger cities (see Supplementary Information for more details).

These two assumptions about complementarity and diversity are enough to generate our results. A wide range of phenomena, including industrial employment, innovation, crime, educational attainment and disease incidence, are all statistically consistent with our theory. Moreover, we reveal an important empirical fact about the factors affecting different urban phenomena: that they change in similar ways across phenomena, implying that all scaling parameters for an urban phenomenon can be obtained from a single observation. This suggests that urban scaling is a highly constrained phenomenon, which in turn allows us to test the theory through its ability to predict the likely prevalence of a phenomenon across cities.

Our work is also related to the literature on production recipes[34], which has recently been applied to explaining performance curves

[1]Center for International Development, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts 02115, USA. [3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA. [4]Harvard Kennedy School, Harvard University, Cambridge, Massachusetts 02138, USA. *e-mail: Andres_Gomez@hks.harvard.edu

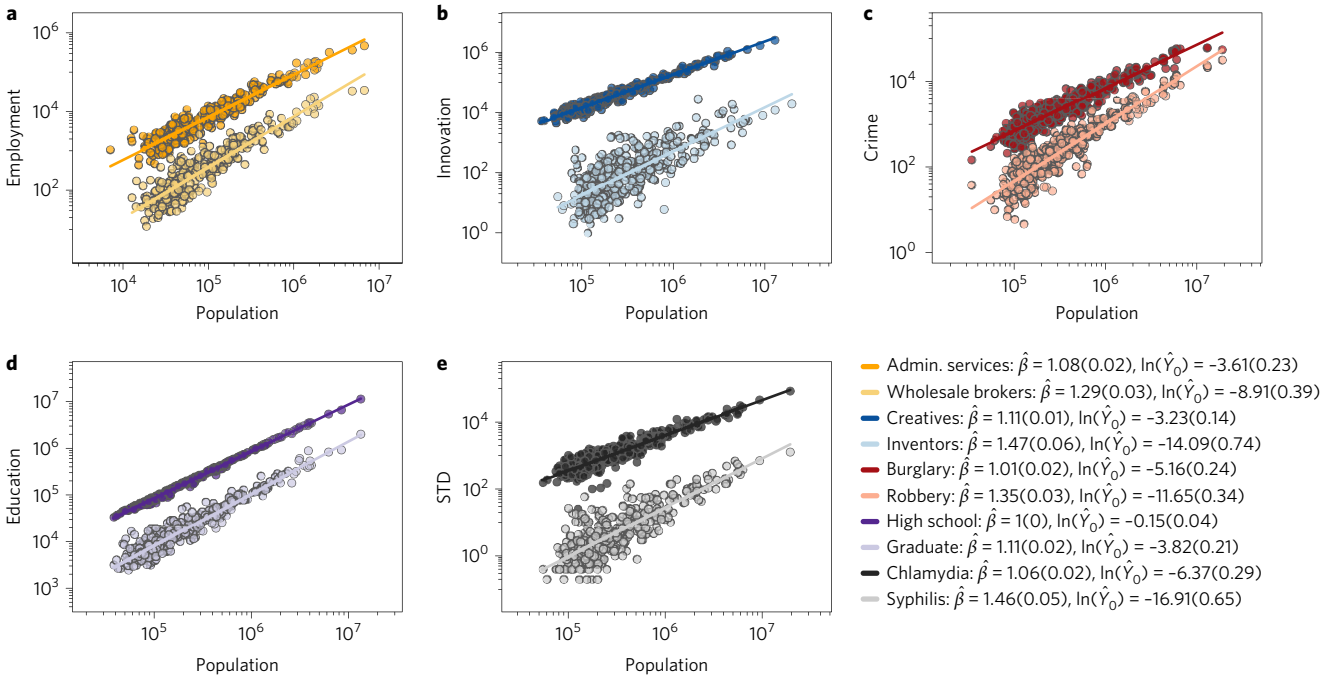


Figure 1 | Four facts across ten different urban phenomena that we seek to explain. Prevalence follows a power-law scaling with population size, and different phenomena have different general prevalence, different scaling exponents, and variance for cities of similar size. **a**–**e**, Scatter plots are shown for the number of individuals in each of the following cases: employment in two industries (**a**), two types of innovative activities (**b**), two types of violent crime (**c**), people with a given educational level (**d**), and two sexually transmitted diseases (STDs) (**e**). Hat (ˆ) denotes a statistical estimate of a parameter. The lines represent the best fit of the model $E\{Y|N\} = Y_0 N^{\beta}$ (see Methods for data sources and additional information).

in production processes[9]. The notion of complementarity, which is central to our approach, also plays a role in the 'componential theory of creativity' by Amabile[35], the 'violentization' model of criminality by Athens[36] and recombinant growth models by Weitzman[37]. The closest approach to our framework, however, is the model of Hausmann and Hidalgo[3], which assumes that industries are present in a location when the elements necessary for the industry are available in that location. They use a simple model in which the number of elements in a location is a binomial random variable with probability r, and the elements required by each industry can be represented by another binomial random variable with probability q. Assuming constant r for all countries and q for all industries, they explain how ubiquitous industries are across countries, the inverse relationship between the diversity of countries and the average ubiquity of their industries, and other relevant statistics. However, they limit the analysis to industry presence and do not look at scaling phenomena. A new conceptual component of our model is that it allows the required factors specific to a given activity to be different for each individual. That is, any two individuals in the population can require two different sets of factors in order to be counted into a given activity.

The parameters of the formal model are listed in Table 1. Each phenomenon has a number of factors M on which it can depend. With probability q an individual requires any one of those M factors, and with probability r a city provides any one of the factors. We model the random variable representing the aggregate output of a given phenomenon as $Y = \sum_{j=1}^{N} X_j$, where $X_j = 1$ if individual j has access to all the required factors she needs in city c to be counted in a given activity, and $X_j = 0$ if she does not, with $j \in \{1, \ldots, N\}$.

Given a city with some factors present in it (from a total of M possible factors), the probability that individual j generates an output (that is, that $X_j = 1$) is the probability that the individual requires none of the factors that the city does not have. Therefore, if an individual is exposed to m factors, the individual must not require any of the other $M - m$ factors that are not present, if his or her output is to be 1. Since the probability that an individual does not require a particular factor is $1 - q$, the probability that an individual is counted in the activity given a city with m factors is $\Pr\{X_j = 1 | M_{\text{city}} = m\} = (1-q)^{M-m}$, where M_{city} is a binomially distributed random variable $\text{Binom}(M, r)$.

It follows that $X_1, \ldots, X_N$ are identically distributed random variables. The expected value of Y is thus $E\{Y\} = N \sum_{m=0}^{M} \Pr\{X_j = 1 | M_{\text{city}} = m\} \Pr\{M_{\text{city}} = m\}$. The variance of Y can be calculated similarly. This yields (see Supplementary Information for the complete derivation):

$$E\{Y\} \approx NP \tag{1}$$

and

$$\text{Var}\{Y\} \approx E\{Y\}^2 \left(\frac{1}{E\{Y\}} - \frac{1}{N} + \frac{1}{P^q} - 1 \right) \tag{2}$$

where $P \equiv e^{-Mq(1-r)}$.

Table 1 | Parameters of the model.

Parameter	Meaning
$N > 0$	City population size susceptible of participating in a given phenomenon
$M > 0$	Number of possible factors required for the given phenomenon
$q \in (0,1)$	Probability that an individual needs any given factor from the environment
$r \in (0,1)$	Probability that the city provides any one of the factors to the individual

The parameters M, q and r are in principle phenomenon-dependent.

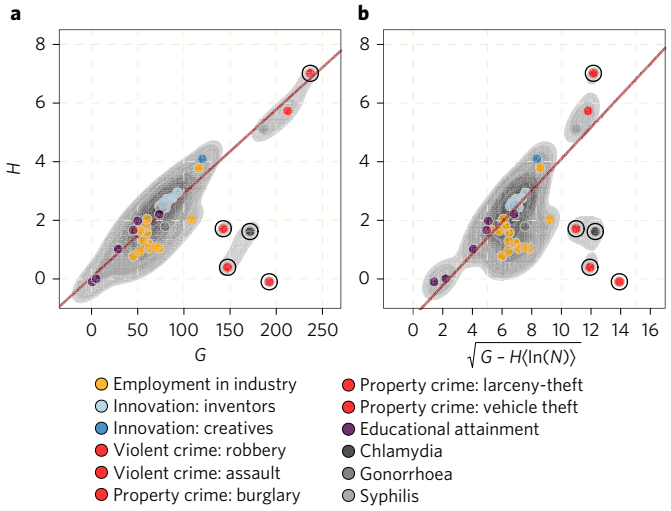


Figure 2 | Relationship between inferred values of parameters *G*, *H* and $\sqrt{G - H\langle \ln N \rangle}$, across 43 different urban phenomena. **a**,**b**, The theory does not constrain their values, so the figure shows in grey the contours of a kernel-density estimate to reveal underlying patterns and relationships. A linear relationship is suggested by the estimated density. The line is the estimated robust regression that excludes the top five outliers ringed in black, which are phenomena with the lowest estimated density. In both panels, the outliers are the same: 'robbery', 'aggravated assault', 'burglary', 'larceny-theft' and 'chlamydia'. The linear trends in both panels are an empirical indication that the coefficients s_1 and s_2 are mostly constant across phenomena.

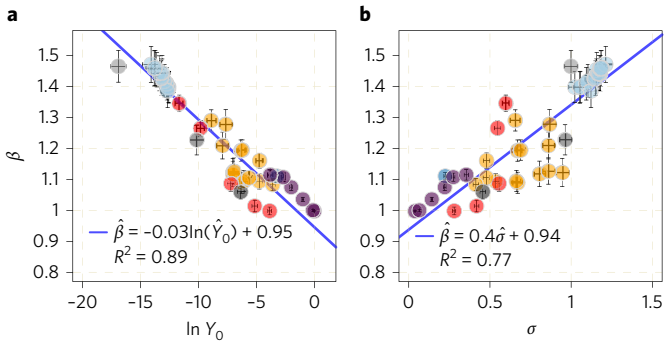


Figure 3 | Predictions. **a**,**b**, The theory predicts a negative linear relationship between β and $\ln Y_0$ (**a**), and a positive relationship between β and σ (**b**), both with an intercept of 1. As a consequence, there is an implied negative linear relationship between σ and $\ln Y_0$ with zero intercept. Both figures show the point estimates (the centres of the coloured disks) and the corresponding standard errors of the estimated parameters of the scaling laws (the error bars) for each of the 43 urban phenomena studied.

Since r is the fraction of factors that an individual is expected to encounter in a city, r represents a measure of urban diversity. This parameter captures the accumulation of factors in the population. As we have argued, factors tend to accumulate logarithmically with population size when a process of selection is involved (see Supplementary Information for more details). Factors can be acquired by individuals through a process of social learning as in models of cultural evolution, or by cities as a whole as they integrate individuals with qualitatively new and different characteristics, skills, behaviours, beliefs, occupations or tools.

We thus assume that $r = a + b\ln(N)$. Replacing r in equation (1) yields the scaling function $\mathrm{E}\{Y\} = Y_0 N^{\beta}$ (see equations (3) and (4) below). Hence, the power-law scaling of phenomena with population size across cities emerges from two relations that offset each other: the exponential relation between the prevalence of a phenomenon in a city and diversity, and the logarithmic relation of diversity to population size. We hypothesize that power-law scaling does not emerge if diversity does not scale logarithmically with population size. In this way, our theory can potentially reconcile observations in which power-law scaling breaks down (for example, for small population sizes[38]) and can also be consistent with other scale-dependent functions, such as $\mathrm{E}\{Y\} = Y_0 N \ln(N/N_0)$ (see refs [39,40]), which can arise if diversity scales more slowly than logarithmically (see ref. [33]). We thus provide theoretical support to a wide empirical literature on urban scaling[1,38,41–43].

Furthermore, our model predicts that the logarithm of the general prevalence of a particular phenomenon, its scaling exponent and the average standard deviation across population sizes all change linearly according to the complexity of the phenomenon (see Supplementary Information for the precise derivation). Since the parameter q is the fraction of factors that an individual is expected to require from the city in order to be counted into a phenomenon, q quantifies the complexity of that phenomenon. Specifically, we have

$$\ln Y_0 = -M(1-a)q \tag{3}$$

$$\beta - 1 = Mbq \tag{4}$$

$$\sigma = \sqrt{M(1-a-b\langle \ln N \rangle)}\, q \tag{5}$$

where $\sigma \equiv \sqrt{\langle \mathrm{Var}\{\ln Y\} \rangle}$, with $\langle \cdot \rangle$ being the mean across population sizes, such that $\ln(N)$ is the mean of the logarithm of population sizes. In short, an increase in the complexity q of a phenomenon (for example a decrease in transmissibility of a disease that makes it more difficult to acquire) would simultaneously decrease the intercept, increase the scaling exponent and increase its variance in cities of same population size. In other words, complex phenomena are expected to be rare and to scale steeply with population size, and their prevalence will be subject to high stochastic variability.

Conditioned on knowing β, $\ln Y_0$ and σ, equations (3), (4) and (5) represent three equations with four unknowns. The equations can then be solved for $G = M(1-a)$, $H = Mb$ and q (leaving M, the total possible number of factors that affect each phenomenon, undetermined).

We estimate β and $\ln Y_0$ through ordinary least squares (OLS), and estimate σ as the square root of the mean squared error of the OLS regression. We then solve for G, H and q. Interestingly, even though G and H vary widely across phenomena, the ratio $s_1 = H/G$ remains numerically stable, as manifested in Fig. 2a in which G and H feature a linear relationship with zero intercept. In this ratio, the parameter M factors out of H and G and cancels, yielding $s_1 = b/(1-a)$. This suggests that the parameters for how diversity changes with population size (a and b) are related in the same way across all phenomena. Similarly, the fact that G is almost two orders of magnitude larger than H signifies that the ratio $s_2 = H/\sqrt{G - H\langle \ln N \rangle}$ also remains approximately stable (Fig. 2b). This is because the ratio goes as $c\sqrt{G}$ with $c \to 0$. These ratios are important because they connect the scaling parameters: namely, $\beta = 1 - s_1 \ln Y_0$ from equations (3) and (4), and $\beta = 1 + s_2\sigma$ from equations (4) and (5). As a consequence, the way in which β changes with a change in $\ln Y_0$ and σ, respectively, is similar across activities. In other words, the implication of Fig. 2 is that we can plot the estimated values of β versus $\ln Y_0$ and β versus σ for different activities in the same graph, and expect them to be linearly related. Figure 3 shows that this is indeed the case. The implication is that these three scaling parameters are strongly constrained in the parameter space and lie in a line.

Provided that the coefficients s_1 and s_2 are constants and are known in advance, the theory therefore establishes that knowing the value of one of the scaling parameters of a phenomenon of

Box 1 | What the theory predicts

From the prevalence of a phenomenon in a single city, the theory predicts what the prevalence in the remaining cities is likely to be.

Procedure

Given coefficients s_1 and s_2 and the populations of all cities $n_1, n_2, \ldots$

For a phenomenon of interest, pick a random city c with known population size and prevalence:

$$(n_c, y_c)$$

Apply the equations (where 'pred.' indicates a predicated value):

$$\beta^{(\text{pred.})} = \frac{1 - s_1 \ln(y_c)}{1 - s_1 \ln(n_c)}$$

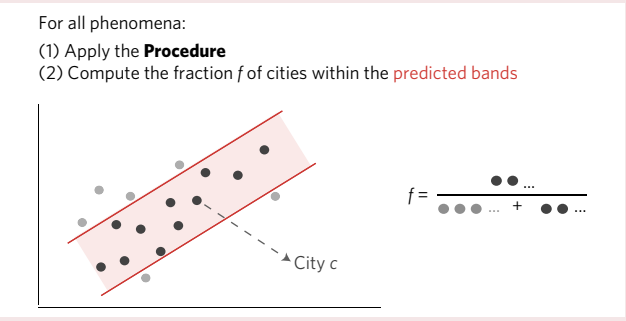
$$\ln\,(Y_0)^{(\text{pred.})} = \frac{1 - \beta^{(\text{pred.})}}{s_1}$$

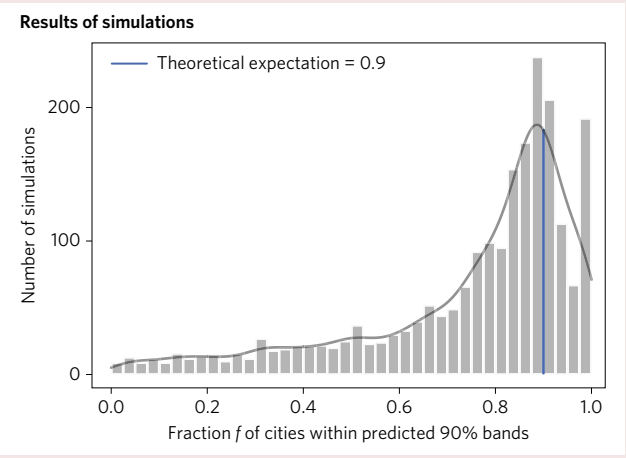
$$\sigma^{(\text{pred.})} = \frac{\beta^{(\text{pred.})} - 1}{s_2}$$

Use the populations n_1, n_2, ..., to predict the prevalence of the phenomenon in the remaining cities within some prediction bands:

$$y_i^{\pm} = \exp\{\ln(Y_0)^{(\text{pred.})} + \beta^{(\text{pred.})} \ln(n_i) \pm z_a \sigma^{(\text{pred.})}\}, \quad \text{for all } i = 1, 2, \ldots$$

To test the predictions, we simulated the procedure 50 times for each phenomenon, for a total of 2,150 simulations.





interest (exponent, general prevalence or variance) determines the value of the others. If unknown, however, this one degree of freedom, in turn, can be fixed if we know the population $N = n_c$ and prevalence $Y = y_c$ in a single city c. This is possible if we assume that the city is an average city, and that the prevalence of the phenomenon is what is expected from its population size, $y_c = Y_0 n_c^{\beta}$. Thus, we can test the theory according to its ability to predict the prevalence of a phenomenon in other cities having knowledge of only one random data point (the prevalence of the phenomenon in a single city). Box 1 explains the step-by-step procedure to determine bands between which the prevalence of a phenomenon is predicted to lie. To test this empirically, we use as an approximation of the median of s_1 and s_2 across phenomena in our dataset, $s_1 \approx 0.03045$ and $s_2 \approx 0.33450$. We pick bands that are $z_{0.95} \approx 1.645$ standard deviations from the mean, so that if the theory is correct, 90% of cities are expected to fall within the bands. For each of the 43 activities in our dataset, we simulate the procedure 50 times, picking a city at random each time (with replacement). The histogram of Box 1 shows the distribution of the fraction of cities f that fall within the bands as a result of the $43 \times 50 = 2{,}150$ simulations.

Here, we are using the proposed prediction framework to test the validity and scope of the theory. But using this framework as an actual tool for predicting the prevalence of a phenomenon in cities where data are unreliable or unavailable is still premature. Further investigations and more data are needed to improve our theory and its practical utility. Moreover, it is important to keep in mind that our results so far imply that more complex phenomena have a higher variability. So even if the theory is correct, 90% prediction bands for complex phenomena can be as wide as two orders of magnitude, and this intrinsic variability affects the practical use of such predictions.

There are two main reasons that some phenomena may deviate from our predictions. First, some of the counts for Y are actually counts over a time period, which may arbitrarily shift the values that $\ln Y_0$ takes, depending on the length of the period. For example, there is no reason why output must be computed as counts per year, as opposed to per month, or something else depending on the activity. And second, the scaling of output, according to the theory, is with respect to the potential population N that is 'susceptible' of engaging in the activity or phenomenon (for example, women, adults or the working age population). Hence, N is not necessarily the whole population of the city, and our estimations of $\ln Y_0$ carry that error from measuring incorrectly the size of the appropriate population group. In spite of these effects, the results in Fig. 3 are broadly consistent with the model.

The theory we present is unabashedly simplified, avoiding issues of supply or demand, equilibrium or the structure of social networks. We have assumed, for example, that people interact with the city as a whole, abstracting away interactions between individuals. We modelled each city as a set of factors, but we did not specify how factors appear. We introduced the notion of the complexity of a phenomenon, representing an average measure of how many inputs an individual needs from the city to be able to be counted, or engage, in the given phenomenon. In the context of epidemiology, we have assumed the diversity of factors necessary for disease transmission to be mostly affected by socio-economic aspects, themselves subject to cultural evolution; similarly with crime. Disease and crime, however, are the subject of strong public policy interventions aimed at reducing their influencing factors. How our model applies to these phenomena is a question that needs to be further analysed as more data are collected.

We have also abstracted away important aspects of cities. First and foremost, we have presented a static view of cities. Also, we have bypassed the interdependencies between cities, and between activities, that arise from people migrating in and out of them[44]. Labour migration and the sharing of resources among cities in a

region can affect the diversity of factors to which a city is exposed and has access. Hence, factors imported from a wider region can affect the prevalence of urban phenomena. Further work is needed on the inclusion of these interactions into the model and their consequence for scaling. We have also left out the dynamic component involved when economic actors act according to complex decision rules. Finally, we have not taken into account the fact that economic and social actors exist not only at the level of individuals but also at intermediate levels of organization such as families, neighbourhoods, firms and so on.

Accordingly, we do not expect the predictions of this model to be numerically accurate, and yet they are reasonable. It is surprising that such a simple model can explain scaling, prevalence and variance of such heterogeneous phenomena in an integrated framework. This indicates that the theory has captured something fundamental about social systems: namely, that they are complex stochastic processes that involve many complementary factors accumulating through evolutionary processes. Models that incorporate these elements can have broad applications in social science.

Methods

Regression analysis. Although our response variables Y are conceptually 'counts', in practice some of our data represent time averages or estimates from statistical offices. We are trying to analyse under a unified framework our data which include both continuous and count variables. For count variables, the use of negative binomial, Poisson or zero-inflated regression analyses is generally preferred over ordinary least squares (OLS), given that the latter assumes a continuous normal conditional distribution of the response and does not allow for the use of zero counts when the regression is done over the logarithm of Y. All of these methods should, in principle, yield similar coefficient estimates to each other, and are rather intended to produce better estimates of their standard errors under different circumstances.

Since our analysis depends on comparing the estimated regression parameters across several urban phenomena, we have opted for the use of OLS regression for all phenomena throughout our analysis. The use of different regression models does not markedly change our estimations, as expected.

Kernel density estimation. In Fig. 2 of the main text, we show the values of G, H and $\sqrt{G-H\langle \ln N\rangle}$ across 43 different urban phenomena. To reveal patterns in the distributions of these values, we applied a two-dimensional kernel density estimation separately for G and H, and $\sqrt{G-H\langle \ln N\rangle}$ and H. See the Supplementary Information for an analysis of the outliers and how they affect the linear relationship.

We used the R package 'ks', freely available online[45], which uses standard normal kernels with a conventional plug-in selector for the matrix bandwidth estimation. A useful feature of this package is that it allows non-zero values for the non-diagonal elements in the matrix.

Data availability. The data sources are explained below and provided as a Supplementary Data zip file. This contains a single file for each urban phenomenon we studied (except for sexually transmitted diseases, which we kept in a separate file), a README file and an Excel file, which lists the different phenomena we used in our analysis with other parameters and field descriptions.

Employees by industry. Data were downloaded using the programming codes that have been made available by the US Bureau of Labor Statistics through the website http://www.bls.gov/cew/doc/access/data_access_examples.htm. The specific data for micropolitan and metropolitan areas were selected using the guide in http://www.bls.gov/cew/doc/titles/area/area_titles.htm.

The metropolitan codes, however, are from the 2004 definitions. In http://www.bls.gov/cew/cewfaq.htm#Q18, it says "QCEW data for Metropolitan Statistical Areas (MSAs) for the years 1990 to present are based on the March 2004 MSA definitions. Aside from a few titling changes, there have been relatively few updates to those definitions since the March 2004 release. The next major revision to MSA definitions is expected in 2013. The QCEW program will release data for 2013 and forward based on those definitions".

These definitions do not match completely. From http://www.bls.gov/news.release/metro.nr0.htm, "The Metropolitan New England City and Town Areas (NECTAs) and NECTA Divisions again are used for the six New England states, rather than the county-based delineations, for purposes of this news release".

The list of industry codes can be found in http://www.bls.gov/cew/doc/titles/industry/industry_titles.htm. The US Bureau of Labor Statistics uses the North American Industry Classification System (NAICS) to assign establishments, and their employment numbers, to different industries. NAICS uses numeric codes to classify establishments at different levels of aggregation, from very detailed to highly aggregated categories of industries. The more detailed industrial categories are described by 6-digit codes, whereas the most aggregated categories use two digits. We aggregate the employment statistics at the 3-digit code classification level. From the 91 different industries, we pick only those industries that have presence (at least one employee) in more than 250 metropolitan areas. This is to ensure that the statistical significance is comparable with the other urban phenomena. Since our theory does not yet account for sublinear scaling of phenomena, we pick the industries that have scaling exponents larger than 1 for employment with population size. This reduces the sample of three-digit industries from 91 to 14. Our results, however, are robust to the inclusion of more (superlinear) industries with presence in less than 250 MSAs.

Sexually transmitted diseases. The data on sexually transmitted diseases (STDs) consist of new cases of chlamydia and syphilis (primary, secondary and congenital). They represent the 5-year cumulative incidence, from 2007 to 2011, in the counties of the 48 contiguous states of the United States, as reported by the Centers for Disease Control and Prevention[46]. In our analysis, we used the average of counts over the years 2007–2011.

The surveillance information in this dataset is based on the following sources of data: (1) notifiable disease reporting from state and local STD programmes; (2) projects that monitor STD positivity and prevalence in various settings, including the National Job Training Program, the STD Surveillance Network, and the Gonococcal Isolate Surveillance Project; and (3) other national surveys implemented by federal and private organizations. This dataset does not include any individual-level information on reported cases.

Since the STD data were originally obtained at the county level, we constructed MSA-level metrics using county-level data (see ref. [43] for details). Of the 375 MSAs within the 48 contiguous states, our dataset has information on 364.

Creative individuals. Here we use the definition of 'creative occupations' given by the US Department of Agriculture (USDA, http://www.ers.usda.gov/data-products/creative-class-county-codes/documentation), as an improvement to that originally proposed by Florida[47]. The USDA defines these occupations: "O*NET, a Bureau of Labor Statistics data set that describes the skills generally used in occupations, was used to identify occupations that involve a high level of 'thinking creatively.' This skill element is defined as 'developing, designing, or creating new applications, ideas, relationships, systems, or products, including artistic contributions.'

The data are available at the county level and have to be aggregated using the 2003 MSA definitions which can be found at http://www.census.gov/population/estimates/metro-city/0312msa.txt. The number of MSAs according to this definition is 361 for the 48 contiguous states. To obtain the MSA populations, we reconstruct them from Census tracks data, aggregating the 2010 populations of counties available at https://www.census.gov/population/metro/data/c2010sr-01patterns.html.

Inventors. Counts of inventors are publicly available through the US Patent and Trademark Office website at http://www.uspto.gov/web/offices/ac/ido/oeip/taf/inv_countyall/usa_invcounty_gd.htm. According to the link (http://www.uspto.gov/web/offices/ac/ido/oeip/taf/reports.htm) this report applies to US resident inventors who have received a utility patent (that is "patent for invention") granted by USPTO since 2000. The report includes a series of tables that display US states and the regional components (for example counties) in which the inventors resided. Counts of the inventors and their patents are provided for each of the regional components.

The documentation can be found in http://www.uspto.gov/web/offices/ac/ido/oeip/taf/inv_countyall/usa_invcounty_gd.htm. In Figs 2 and 3, we plot the years 2000 to 2013 using the 2013 definition of Metropolitan Statistical Areas in terms of counties according to the US Census Bureau (see https://www.census.gov/popest/data/metro/totals/2013/CBSA-EST2013-alldata.html). We merged to this dataset the MSA populations, from 2000 to 2013, reported by the Bureau of Economic Analysis.

Crime. Data for different types of crimes at the MSA level are collected by the Federal Bureau of Investigation (FBI). These data are publicly available at the official website https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/ for different years. In our study, we limit our analysis to the years 2010, 2011 and 2012.

Two important caveats are needed about the crime statistics used in our analysis. On the one hand, we would ideally like to have counts over some period of time of unique individuals who were victims of each different type of crime (we would also like to have counts of criminals in urban areas, but this is obviously difficult to measure). We have proxied the number of victims by the counts of crimes. On the other hand, our model provides predictions for counts of people Y who engage in a given activity, and we compare these counts with the population N that is susceptible to this activity. For most activities, N is easy to define and is typically the total population size of a city. For other activities, N is not so easy to define. Hence, we have removed from our analysis (see Figs 2 and 3) 'murder and nonnegligent manslaughter', and 'forcible rape', defined by the FBI

(see https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/violent-crime/rapemain) as "the carnal knowledge of a female forcibly and against her will". The relevant population N that corresponds to these types of violent crimes is not the total population size of a city but represents a restricted part of the total population, and we think that these phenomena require analysis that is beyond the scope of our model. Misspecifications of N in our regression. produce a bias in the estimation of ln Y_0. We avoid such misspecifications by removing these two types of crimes from our analysis.

Educational attainment. We have used the estimates of the population by the different types of educational attainment from the 2009–2013, 5-Year American Community Survey (ACS) from the US Census Bureau. We have used as the base population N the population of 25 years and older.

This dataset is accessible through the website American FactFinder, at http://factfinder.census.gov/ by selecting Advanced Search and entering S1501 as the topic, corresponding to educational attainment. We selected the 5-Year ACS data for 2009–2013, and in 'Geographies' we selected data for all US Metropolitan Statistical Areas.

We adjusted the educational attainment categories to reflect increases in complexity. Hence, from least to most complex, we defined six categories (1) ninth grade, or higher; (2) high school graduate, or higher; (3) some college, or higher; (4) associate's degree, or higher; (5) bachelor's degree, or higher; and (6) graduate or professional degree.

Received 31 May 2016; accepted 1 November2016; published 22 December 2016

References

1. Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl Acad. Sci. USA* **104,** 7301–7306 (2007).
2. Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proc. Natl Acad. Sci. USA* **106,** 10570–10575 (2009).
3. Hausmann, R. & Hidalgo, C. A. The network structure of economic ouput. *J. Econ. Growth* **16,** 309–342 (2011).
4. Henrich, J. Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses—the Tasmanian case. *Am. Antiq.* **69,** 197–214 (2004).
5. Schroeder, M. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* (Freeman, 1991; republished Dover, 2009).
6. Sornette, D. *Critical Phenomena in Natural Sciences—Chaos, Fractals, Selforganization and Disorder: Concepts and Tools* 2nd edn (Springer, 2006).
7. West, G. B. & Brown, J. H. The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. *J. Exp. Biol.* **208,** 1575–1592 (2005).
8. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **453,** 779–782 (2008).
9. McNerney, J., Farmer, J. D., Redner, S. & Trancik, J. E. Role of design complexity in technology improvement. *Proc. Natl Acad. Sci. USA* **108,** 9008–9013 (2011).
10. Batty, M. The size, scale, and shape of cities. *Science* **319,** 769 (2008).
11. West, G. B., Brown, J. H. & Enquist, B. J. A general model for the origin of allometric scaling laws in biology. *Science* **276,** 122–126 (1997).
12. Banavar, J. R., Maritan, A. & Rinaldo, A. Size and form in efficient transportation networks. *Nature* **399,** 130–132 (1999).
13. Arbesman, S., Kleinberg, J. M. & Strogatz, S. H. Superlinear scaling for innovation in cities. *Phys. Rev. E* **79,** 016115 (2009).
14. Pan, W., Ghoshal, G., Krumme, C., Cebrian, M. & Pentland, A. Urban characteristics attributable to density-driven tie formation. *Nat. Commun.* **4,** 1961 (2013).
15. Bettencourt, L. M. A. The origins of scaling in cities. *Science* **340,** 1438 (2013).
16. Yakubo, K., Saijo, Y. & Korošak, D. Superlinear and sublinear urban scaling in geographical networks modeling cities. *Phys. Rev. E* **90,** 022803 (2014).
17. Banavar, J. R. *et al.* A general basis for quarter-power scaling in animals. *Proc. Natl Acad. Sci. USA* **107,** 15816–15820 (2010).
18. McMahon, T. Size and shape in biology. *Science* **179,** 1201–1204 (1973).
19. West, G. B., Brown, J. H. & Enquist, B. J. The fourth dimension of life: fractal geometry and allometric scaling of organisms. *Science* **284,** 1677–1679 (1999).
20. Samaniego, H. & Moses, M. E. Cities as organisms: allometric scaling of urban road networks. *J. Transp. Land Use* **1,** 21–39 (2008).
21. Hidalgo, C. A., Klinger, B., Barabasi, A.-L. & Hausmann, R. The product space conditions the development of nations. *Science* **317,** 482–487 (2007).
22. Klimek, P., Hausmann, R. & Thurner, S. Empirical confirmation of creative destruction from world trade data. *PLoS ONE* **7,** 1–9 (2012).
23. Henrich, J. & Boyd, R. On modeling cognition and culture: why cultural evolution does not require replication of representations. *J. Cogn. Culture* **2,** 87–112 (2002).
24. Powell, A., Shennan, S. & Thomas, M. G. Late Pleistocene demography and the appearance of modern human behavior. *Science* **324,** 1298–1301 (2009).
25. Kline, M. A. & Boyd, R. Population size predicts technological complexity in oceania. *Proc. R. Soc. Lond. B* **277,** 2559–2564 (2010).
26. Mesoudi, A. Variable cultural acquisition costs constrain cumulative cultural evolution. *PLoS ONE* **6,** e18239 (2011).
27. Derex, M., Beugin, M.-P., Godelle, B. & Raymond, M. Experimental evidence for the influence of group size on cultural complexity. *Nature* **503,** 389–391 (2013).
28. Kempe, M. & Mesoudi, A. An experimental demonstration of the effect of group size on cultural accumulation. *Evol. Hum. Behav.* **35,** 285–290 (2014).
29. Collard, M., Ruttle, A., Buchanan, B. & OBrien, M. J. Population size and cultural evolution in nonindustrial food-producing societies. *PLoS ONE* **8,** e72628 (2013).
30. Bromham, L., Hua, X., Fitzpatrick, T. G. & Greenhill, S. J. Rate of language evolution is affected by population size. *Proc. Natl Acad. Sci. USA* **112,** 2097–2102 (2015).
31. Brummitt, C. D., Gomez-Lievano, A., Goudemand, N. & Haslam, G. Hunting for keys to innovation: the diversity and mixing of occupations do not explain a city's patent and economic productivity. In *Proc. Complex Systems Summer School Santa Fe Institute* 1–13 (2012); https://www.santafe.edu/engage/learn/resources/csss-2012-proceedings
32. Youn, H. *et al.* Scaling and universality in urban economic diversification. *J. R. Soc. Interf.* **13,** http://dx.doi.org/10.1098/rsif.2015.0937 (2016).
33. Bettencourt, L. M., Samaniego, H. & Youn, H. Professional diversity and the productivity of cities. *Sci. Rep.* **4,** 5393 (2014).
34. Auerswald, P., Kauffman, S., Lobo, J. & Shell, K. The production recipes approach to modeling technological innovation: an application to learning by doing. *J. Econ. Dynam. Control* **24,** 389–450 (2000).
35. Amabile, T. *Creativity in Context* (Westview, 1996).
36. Athens, L. H. *The Creation of Dangerous Violent Criminals* (Univ. Illinois Press, 1992).
37. Weitzman, M. L. Recombinant growth. *Q. J. Econ.* **113,** 331–360 (1998).
38. Gomez-Lievano, A., Youn, H. & Bettencourt, L. M. A. The statistics of urban scaling and their connection to Zipf's Law. *PLoS ONE* **7,** e40393 (2012).
39. Shalizi, C. R. Scaling and hierarchy in urban economies. Preprint at http://arxiv.org/abs/1102.4101 (2011).
40. Bettencourt, L. M. A., Lobo, J. & Youn, H. The hypothesis of urban scaling: formalization, implications and challenges. Preprint at http://arxiv.org/abs/1301.5919v1 (2013).
41. Mantovani, M. C., Ribeiro, H. V., Lenzi, E. K., Picoli, S. & Mendes, R. S. Engagement in the electoral processes: scaling laws and the role of political positions. *Phys. Rev. E* **88,** 024802 (2013).
42. Arcaute, E. *et al.* Constructing cities, deconstructing scaling laws. *J. R. Soc. Interf.* **12,** 20140745 (2014).
43. Patterson-Lomba, O., Goldstein, E., Gómez-Liévano, A., Castillo-Chavez, C. & Towers, S. Per capita incidence of sexually transmitted infections increases systematically with urban population size: a cross-sectional study. *Sex. Transm. Infect.* **91,** 610–614 (2015).
44. Neffke, F. & Henning, M. Skill relatedness and firm diversification. *Strateg. Manag. J.* **34,** 297–316 (2013).
45. Duong, T. ks: kernel density estimation and kernel discriminant analysis for multivariate data in R. *J. Stat. Softw.* **21,** 1–16 (2007).
46. Centers for Disease Control and Prevention. *Sexually Transmitted Disease Surveillance 2012* (US Department of Health and Human Services, 2013).
47. Florida, R. *The Rise of the Creative Class: And How It's Transforming Work, Leisure, Community and Everyday Life* (Basic Books, 2004).

Acknowledgements

We thank A.-L. Barabasi, J. Lobo, L. M. A. Bettencourt, F. Neffke, S. Valverde, D. Diodato and C. Brummitt for their comments on this work. We also thank M. Akmanalp and W. Strimling for their suggestions about aesthetics. This work was funded by the MasterCard Center for Inclusive Growth, and Alejandro Santo Domingo. O.P-L. acknowledges support by National Institutes of Health (NIH) grant T32AI007358-26. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

A.G-L. and O.P-L. collected the data, and conceived and designed the study. A.G-L. conducted the analyses. A.G-L. and R.H. developed the model. A.G-L., O.P-L. and R.H. wrote the manuscript. All three authors reviewed and approved the paper.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.G-L.

How to cite this article: Gomez-Lievano, A., Patterson-Lomba, O. & Hausmann, R. Explaining the prevalence, scaling and variance of urban phenomena. *Nat. Hum. Behav.* **1,** 0012 (2016).

Competing interests

The authors declare no competing interests.