Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities

Yannick Malevergne,[1,2,3] Vladilen Pisarenko,[4] and Didier Sornette[3,5]

[1]*Université de Lyon-Université de Saint-Etienne, Coactis E.A. 4161, 42023 Saint-Etienne, France*

[2]*EMLYON Business School, Cefra, 69134 Ecully, France*

[3]*Department of Management, Technology, and Economics, ETH Zurich, Switzerland*

[4]*International Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Science, Moscow, Russia*

[5]*Swiss Finance Institute, c/o University of Geneva, 40 blvd. Du Pont d'Arve CH-1211 Geneva 4, Switzerland*

(Received 9 September 2010; revised manuscript received 20 January 2011; published 22 March 2011)

Fat-tail distributions of sizes abound in natural, physical, economic, and social systems. The lognormal and the power laws have historically competed for recognition with sometimes closely related generating processes and hard-to-distinguish tail properties. This state-of-affair is illustrated with the debate between Eeckhout [Amer. Econ. Rev. **94** , 1429 (2004)] and Levy [Amer. Econ. Rev. **99** , 1672 (2009)] on the validity of Zipf's law for US city sizes. By using a uniformly most powerful unbiased (UMPU) test between the lognormal and the power-laws, we show that conclusive results can be achieved to end this debate. We advocate the UMPU test as a systematic tool to address similar controversies in the literature of many disciplines involving power laws, scaling, "fat" or "heavy" tails. In order to demonstrate that our procedure works for data sets other than the US city size distribution, we also briefly present the results obtained for the power-law tail of the distribution of personal identity (ID) losses, which constitute one of the major emergent risks at the interface between cyberspace and reality.

DOI: 10.1103/PhysRevE.83.036111 PACS number(s): 89.75.Da, 87.23.Ge, 89.20.−a

I. INTRODUCTION

Probability distribution functions with a power-law dependence in terms of event or object sizes seem to be ubiquitous statistical features of natural and social systems [1]. It has repeatedly been argued that such an observation relies on an underlying self-organizing mechanism, and therefore power-laws should be considered as the statistical imprints of complex systems. It is often claimed that the observation of a power-law relation in data often points to specific kinds of mechanisms at its origin, that can often suggest a deep connection with other, seemingly unrelated systems. In complex systems, the appearance of power-law distributions is often thought to be the signature of hierarchy and robustness. In the last two decades, such claims have been made for instance for earthquakes, weather, and climate changes, solar flares, the fossil record, and many other systems, to promote the relevance of self-organized criticality as an underlying mechanism for the organization of complex systems [2]. This claim is often unwarranted as there are many non-self-organizing mechanisms producing power-law distributions [3–6].

Research on the origins of power-law relations, and efforts to observe and validate them in the real world, is extremely active in many fields of modern science, including physics, geophysics, biology, medical sciences, computer science, linguistics, sociology, economics, and more. The present paper contributes to the literature by proposing a methodology to distinguish power-laws from a closely associated family, the lognormal distribution. Indeed, contrary to what the extensive literature would have us to believe, qualifying the tail of a distribution as being a power-law is full of difficulties and traps, leading to many incorrect claims.

Entering a heated debate on the nature of the distribution of city sizes in the US, we show how a specific test can go a long way toward improving the methodology to qualify power-laws. This statistical tool, the uniformly most powerful unbiased (UMPU) test, is shown to provide a clear diagnostic, allowing us to distinguish between the power-law and the lognormal hypothesis, even when the data set is quite small. This method should play a growing role in many fields plagued by similar curses of undersampled tails.

There has been a recent surge in interest in the size distribution of cities and firms and particularly in the exact shape of the upper tail of this distribution and the implications thereof. Many empirical studies as well as theoretical works have provided evidence and support in favor of a power-law distribution of sizes with a tail index close to one, i.e., Zipf's law [7–11]. However, some other recent works suggest that the size distributions could be close to, or evolve toward, the lognormal law [12–14][1] in accordance with the pure Gibrat principle of proportional growth [15].

Determining the exact shape of the tail of the distribution of the sizes of economic entities, such as cities or firms, is of general interest for several reasons. First, as we recall below, the shape can inform on the mechanisms and generating processes of growth [16,17]. The two large classes of theoretical models of the growth dynamics of cities, purely multiplicative or multiplicative with an additive term, can only be distinguished in their prediction for the tail of the distribution of city sizes, as recalled below. Second, the shape is necessary for many socioeconomical problems. It impacts aggregate economic outcomes [18] and financial

[1]In fact Cabral and Mata [12] consider an even broader model based upon an extended generalized γ distribution of the log size of firms which encompasses the normal distribution as a special case. In such a model, when the size distribution of firms departs from the lognormal, it follows an exponentially dampened power-law.

policies on macroeconomic outputs and behavior. There is also the simple quantitative fact for the importance of the tail: the tiny fraction of cities in the tail that is found to be described better by the power-law, as shown below, account for more than 50% of the whole population. Finally, specifically comparing the lognormal to the power-law model, as often done in the literature and discussed below, has important statistical consequences since the former (respectively, latter) distribution has all its statistical moments finite (respectively, only a finite number of moments), which are characteristics of qualitatively different variability. In particular, for the popular Zipf's law (corresponding to the Pareto distribution with exponent equal to 1), standard statistical tools such as sample mean and sample deviation are not applicable (the Law of Large Numbers is invalid), whereas they are fully justified for the lognormal law.

We stress that, so far, most researchers have essentially considered only these two alternatives (lognormal and Pareto). Past authors have provided many justifications for these two competing models, some of which are recalled in the section below on the generation process. Of course, if another competitor appears, it can be compared with the two main contenders using the statistical methods used here.

To provide an illustrative example of this endless debate between the proponents of Zipf's law and those of the lognormal distribution, let us mention that, based upon the US Census 2000 data, Eeckhout [13] reported that the whole size distribution of cities is lognormal rather than Pareto. This conclusion was obtained by use of the Lilliefors test (or L-test) [19,20] for normal distributions applied to the log-sizes of the cities. It is consistent with Gibrat's law of proportionate effect and is rationalized by an equilibrium theory of local externalities in which the driving force is a random productivity process of local economies and the perfect mobility of workers.

In a comment on this article, Levy [10] argues that the top 0.6% of the largest cities of the US Census 2000 data sample, which account for more than 23% of the population, dramatically depart from the lognormal distribution and is more in agreement with a Pareto distribution. The bulk of the distribution actually follows a lognormal but, due to the departure in the upper tail, a χ^2 test unequivocally rejects the null of a lognormal for the largest cities. The nonrejection of the lognormal by the L-test used in [13] is ascribed to the fact that the relative number of cities in the upper tail is very small (only 0.6% of the sample), and the L-test is dominated by the center of the distribution rather than by its tail, where the interesting action occurs. In reply to this comment, Eeckhout [14] provides the 95%-confidence bands of the lognormal estimates based upon the L-test and shows that the tail of the sample distribution of log size is well within the confidence bands, as shown in Fig. 1. The Appendix describes the Kolmogorov and Lilliefors tests.

The origin of the disagreement between Eeckhout and Levy as well as, more generally, between the supporters of each of the two models, can be traced back to the following reason. In statistical testing, one never proves "truth.' One cannot prove that an hypothesis H is "right" or "correct." One either rejects or fail to reject H. The failure to reject H is not a proof that H is the right model. Eeckhouts' hypothesis H is the lognormal

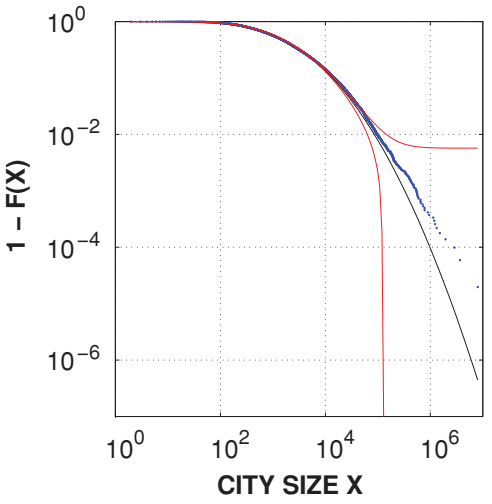

FIG. 1. (Color online) Non-normalized empirical complementary cumulative distribution of city sizes of the US Census 2000 as a function of city sizes (decimal log-log scale). The continuous (black) line is the best fit to the lognormal distribution over the whole city data set. The two (red) lines which fan out strongly in the upper tail delineate the confidence bands generated by the Lilliefors test with 5% significance level, a two-sided goodness-of-fit test suitable when a fully specified null distribution is unknown and its parameters must be estimated. Note that the Lilliefors test statistic is the same as for the Kolmogorov test, for the specific case of testing the fit of the data of logarithm of city sizes with the normal distribution with sample mean and variance. Eeckhout [14] chose a confidence interval of 5% to show that the tail of the distribution is well within this "tight" interval and to conclude incorrectly (see text) that the lognormal hypothesis is not rejected. Regeneration by the present authors of figure 2 of Ref. [14]. See the Appendix.

distribution (LN) on the whole range of city sizes. Levy suggested a more detailed and more general hypothesis H^* that includes H as a particular case. Specifically, H^* implies that city sizes are LN distributed up to some threshold u, and power-law (PL) distributed above this threshold. For u chosen to exceed the maximum city size, evidently H^* coincides with H, i.e., H is a particular case of H^*. Eeckhouts' hypothesis included two unknown parameters: the mean and variance of the logarithm of the city sizes, whereas Levy's hypothesis includes two additional parameters: the threshold u and the Pareto index α (thus, making a total of four parameters). Levy claimed that the uppermost tail (observations exceeding some threshold u) is distributed as a PL. But his conclusions were not universally accepted (and, in particular, they were rejected by Eeckhout), because Levy did not use the most optimal statistical tests in his derivation. As a consequence, statistical scatters of the uppermost observations could be suspected as a possible cause for the deviation of the empirical tail from the LN law and for its visual resemblance to the PL. At this stage, we should already note a problem in the logic of the rejection by Eeckhout [14] of Levy's arguments. Indeed, the Appendix shows that Eeckhout was not correct when he used the Lilliefors test to support the null hypothesis that the distribution of city sizes is lognormal in the tail range, while failing to mention that the Lilliefors test at the same time rejects the null hypothesis for the whole range of city sizes.

Here, in order to provide a definitive answer and to close the debate about the shape of the upper tail of the size distributions of cities, we use the uniformly most powerful unbiased (UMPU) test of the null hypothesis that the upper part of the size distribution (exceeding some threshold u) is a power-law, against the alternative hypothesis that it follows a (truncated from below) lognormal law. We used the maximum likelihood method to estimate the appropriate threshold u, separating the LN from the PL and found $u \simeq 37\,000$ inhabitants. There are about 1000 largest cities above this threshold, and they contain more than 50% of the total city dwellers in the United States. Thus, we unambiguously conclude that the distribution of the 1000 largest US cities follows a power-law. As to the lower part of city size distribution (i.e., the cities smaller than the threshold u), the truncated (from above) lognormal law is an excellent model.

This article is organized as follows. We summarize in Sec. II the properties that often make difficult the task of distinguishing between the Pareto and the lognormal distributions. While the Pareto and the lognormal distributions have indeed distinct asymptotic tails—in contrast with the Pareto, the lognormal *is not* regularly varying but rapidly varying—the lognormal can easily be mistaken for a Pareto over a range which can cover several decades as soon as its standard deviation is sufficiently large (a few units is sufficient). Furthermore, both distributions may be generated by Gibrat's law of proportional growth, with some additional apparently innocuous but actually profound twist(s) for the Pareto. In Sec. III, we use the uniformly most powerful unbiased test of the Pareto distribution against the lognormal. It enables us to find one reason for the disagreement between Eeckhout and Levy, as resulting from the limited power of their tests. More generally, using this uniformly most powerful unbiased test, we confirm and extend Levy's result, by showing that the Pareto model holds for the 1000 largest US cities or so, i.e., for more than 50% of the total human population and that most firm sizes across the world are also distributed according to a power-law. However, Zipf's law, corresponding to Pareto with exponent 1, is found incompatible with most of the data samples. The Pareto index for the uppermost tail (about 1000 largest cities) is approximately 1.4.

II. WHY THE PARETO AND THE LOGNORMAL DISTRIBUTIONS ARE DIFFICULT TO DISTINGUISH

A. Structural similarities and differences

In order to justify that Levy's results are compatible with his own, Eeckhout [14] asserts that both the Pareto distribution and the lognormal distribution are regularly varying, which makes their tail indistinguishable. We recall that a positive function $f(x)$ is regularly varying at infinity if there exists a finite real number α such that [21]

$$\lim_{x\to\infty} \frac{f(t\cdot x)}{f(x)} = t^{\alpha}, \quad \forall t > 0. \tag{1}$$

Pareto distributions are regularly varying. However, it is not the case for lognormal distributions. Indeed, the lognormal density reads

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\cdot\frac{1}{x}\,e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \tag{2}$$

so that

$$\lim_{x\to\infty}\frac{f(t\cdot x)}{f(x)} = \lim_{x\to\infty}\frac{1}{t}\,e^{-\frac{(\ln t)^2}{2\sigma^2}}e^{-\ln t\cdot\frac{\ln x-\mu}{\sigma^2}} = \begin{cases} 0, & t>1,\\ 1, & t=1,\\ \infty, & t<1. \end{cases} \tag{3}$$

This limit behavior characterizes a rapidly decreasing function at infinity. Therefore, Pareto and lognormal distributions exhibit *qualitatively* different behaviors in their upper tails. The lognormal density goes to zero faster than any Pareto density. In this respect, they cannot be mistaken into one another, provided that one has enough data to sample the tail.

However, writing the lognormal density as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\cdot\frac{1}{x}\,e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\mu^2}{2\sigma^2}}\cdot x^{-1+\frac{\mu}{\sigma^2}-\frac{\ln x}{2\sigma^2}}, \tag{4}$$

we observe that the lognormal distribution is superficially like a Pareto distribution with a slowly increasing effective exponent

$$\alpha(x) = \frac{1}{2\sigma^2}\ln\left(\frac{x}{e^{2\mu}}\right). \tag{5}$$

Expression (5) allows us to make two points. First, as stated above, it shows that the lognormal distribution decays at infinity faster than any Pareto distribution, since the apparent exponent $\alpha(x)$ diverges with x. Second, if σ^2 is large enough, the apparent exponent $\alpha(x)$ varies so slowly so as to give the impression of constancy over several decades in x. Quantitatively, in the range $X \leqslant x \leqslant \lambda X$, the apparent exponent varies from $\alpha(X)$ to $\alpha(X) + \frac{1}{2\sigma^2}\ln\lambda$. For instance, for $\sigma = 3.4$, the apparent exponent varies by no more than 0.3 over three decades ($\lambda = 1000$), as illustrated in Fig. 2.

In the case of the US Census 2000 data, with the smaller estimate $\hat{\sigma} = 1.25$ provided in [13], the apparent exponent varies by 1.5 units over just two decades. This is an indication that a powerful test should be able to distinguish the two hypotheses over a range of two to three decades corresponding to the tail regime.

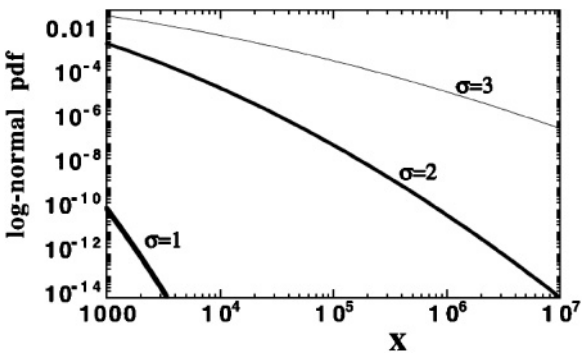

FIG. 2. The lognormal probability density function with $\sigma = 2$ and 3 are close to linear over almost four decades both in abscissa and ordinate in this log-log plot, in which an exact straight line qualifies a power law distribution. With some additional noise, it would be difficult to distinguish them from pure power-laws with constant exponents.

B. Generating process

Gibrat's law of proportional growth is often taken as a key starting point to understand the origin of the distribution of city sizes (see the recent review [22] and references therein). Considered as the unique ingredient, Gibrat's law predicts that the distribution of city sizes should tend to a lognormal distribution, but as a more and more degenerate one as time increases (here, the vocable "degenerate" refers to the fact that all the realizations shrink to zero asymptotically). Indeed, Gibrat's law leads to model the growth of a given city as following a random walk in its log size, which therefore never admits a steady state distribution. Let us also mention [23], that introduced a general class of self-similar fragmentation processes generalizing Gibrat's law, that converges in distribution to the lognormal law, which therefore appears as a robust attractor of a large class of processes.

The equation of city/firm growth embodying Gibrat's law is

$$S_{i,t} = a_{i,t} \cdot S_{i,t-1}, \tag{6}$$

where $S_{i,t}$ is the size of city/firm i at time t and $a_{i,t}$ is the random positive growth factor. Taking the logarithm of Eq. (6) and iterating yields

$$\ln S_{i,t} = \ln S_{i,t-1} + \eta_{i,t} = \ln S_{i,0} + \eta_{i,1} + \eta_{i,2} + \cdots + \eta_{i,t}, \tag{7}$$

where $\eta_{i,t} \equiv \ln a_{i,t}$. Assuming (for a time) that terms $\eta_{i,t}$ are i.i.d. (independent identically distributed) random variables with expectation A and standard deviation B, the Central Limit Theorem of Probability Theory gives

$$\ln S_{i,t} \simeq t \cdot A + t^{1/2} B \cdot \xi, \tag{8}$$

where ξ is a standard Gaussian random variable $N(0,1)$. Of course, the stationarity of the $\eta_{i,t}$'s should be verified by an appropriate analysis. Assuming in addition that the stochastic growth process for a typical city as a function of time is equivalent to sampling the growth of many cities at a given instant, i.e., that a strong form of ergodicity holds, expression (8) ensures that the distribution of city sizes is lognormal, i.e., the variable $\frac{\ln S_{i,t} - t \cdot A}{t^{1/2} B}$ is $N(0,1)$.

An apparently minor modification of Gibrat's law (6) leads to a bona fide steady state and, therefore, to a stationary distribution. An example of such a minor modification, among many other forms [24], consists in adding a small positive random term $\varepsilon_{i,t}$ to the right-hand side of Eq. (6):

$$S_{i,t} = a_{i,t} \cdot S_{i,t-1} + \varepsilon_{i,t}, \tag{9}$$

where the factors $a_{i,t}$ are, as earlier, positive random factors. The term $\varepsilon_{i,t} > 0$ prevents the small cities from becoming too small and degenerate. In absence of $\varepsilon_{i,t}$, expression (9) is nothing but the random walk in log size leading to the lognormal distribution obtained from Eq. (8). When all the factors $a_{i,t}$ are taken equal to a constant a, Eq. (9) reduces to the well-known autoregression process. The necessary condition for stationarity of the autoregression process is $|a| < 1$. When the factors $a_{i,t}$ are allowed to become random, with sometimes values larger than 1, it is not guaranteed that a stationary distribution for $S_{i,t}$ exists even when the series of terms $\varepsilon_{i,t}$ is stationary. Kesten [25] derived the necessary condition for the stationarity of the process (9), which reads $\mathrm{E}[\ln a_{i,t}] < 0$. Roughly speaking, this means that the "growth rates" $\ln a_{i,t}$ of the multiplicative factors should have a negative bias to prevent the divergence of the process (9). Moreover, it is proven that the limit distribution of Eq. (9) has a PL tail with index α that is the strictly positive solution of the equation $\mathrm{E}[(a_{i,t})^{\alpha}] = 1$. The presence of the "minor excitation term" $\varepsilon_{i,t}$ ensures that the size distribution of cities switches for large values from a lognormal to a Pareto law. Because the process (9) with nonzero $\varepsilon_{i,t}$ leads to a stationary distribution, if we assume ergodicity, then the distribution of an ensemble of cities at a give time is the same as that of the set of realizations $\{S_{i,t}\}$ for a fixed city i as a function of t for large times. Gabaix [9] argued for the validity of the constraint $\mathrm{E}[a_{i,t}] = 1$, which then leads automatically to Zipf's law ($\alpha = 1$), but it seems that this restriction is questionable. Zipf's law is obtained more generally as a result of Gibrat's law for large sizes holding together with a condition balancing the birth rate, random growth, and possible death rate of cities.[2] While Zipf's law is quite fashionable, there are often departures from the value $\alpha = 1$. In fact, as we show below, the power index in the PL-tail of city sizes is found close to 1.4.

The intuition behind the transformation of the lognormal into the Pareto distribution, upon the introduction of the apparently minor additive term $\varepsilon_{i,t} > 0$ is the following. Because of the stationarity condition $\mathrm{E}[\ln a_{i,t}] < 0$, in the absence of $\varepsilon_{i,t}$, the process $S_{i,t}$ tends to shrink stochastically toward zero, while exhibiting a more and more degenerate lognormal distribution (here, the vocable "degenerate" refers to the fact that all the realizations shrink to zero asymptotically). During this phase, a few excursions of exponentially large sizes associated with transient occurrences of the growth factor $a_{i,t}$ larger than 1 can occur with exponentially small probability. The term $\varepsilon_{i,t}$ allows the process to repeatedly exhibit the exponentially rare exponentially large excursions. The combination of these two exponentials leads to the Pareto distribution.[3] In sum, reinjection by $\varepsilon_{i,t}$ and transient explosive growth are the two key ingredients for the transformation of lognormal into Pareto in this model (9).

As we have recalled, Gibrat's law of proportional growth yields lognormal distributions, while simple modifications of Gibrat's law for small sizes lead to Pareto distributions. In order to understand the implications of our finding that the tail of the distribution of city sizes is Pareto, we should stress that the statement "Gibrat's law of proportional growth yields lognormal distributions" has to be complemented by the remark that the lognormal distributions are not stable, in the sense that, as the time increases and the system evolves, the mode and mean of the lognormal distributions either converge to zero or diverge to infinity. In contrast, the modified

[2] In the case of cities, death means falling below a moving threshold for qualifying as a city.

[3] For the more realistic situation where cities are on average growing, by an exponentially growing term $\varepsilon_{i,t}$ so as to represent immigration or population fluxes across cities for instance, the same reasoning applies once a change of frame has been performed with respect to the exponentially growing $\varepsilon_{i,t}$ term (see [24] for details).

Gibrat's law for small sizes makes the dynamics stationary when it corresponded to a lognormal distribution converging to zero when the pure Gibrat law holds. In other words, the modification of Gibrat's law for small sizes leading to the Pareto distribution invades all sizes.

III. TESTING THE PARETO AGAINST THE LOGNORMAL DISTRIBUTION

A. Preliminary considerations on statistical testing and UMPU test

Some comments on statistical testing and on the corresponding terminology are relevant to help introduce the "uniformly most powerful unbiased test" used in the present paper.

In statistical testing procedures, one typically considers a sample of i.i.d. (independent identically distributed) random values $(x_1,x_2,\ldots,x_n)$, where the x_k's have the probability density function (PDF) $f(x|\theta)$ that depends on parameter θ (θ may be a vector) belonging to some parametric space Ω. Due to the i.i.d. properties, the PDF of the whole sample is the product $f(x_1|\theta)f(x_2|\theta)\ldots f(x_n|\theta)$. There are two alternative hypotheses on the particular value θ that parameterizes the PDF $f(x|\theta)$ describing our sample:

(1) hypothesis H that parameter θ belongs to some subset Ω_H, or

(2) hypothesis K that parameter θ belongs to Ω_K, which is the complement of Ω_H, i.e., $\Omega_K = \Omega\backslash\Omega_H$.

For instance, $H : \theta = 0; K : \theta > 0$. In this case, Ω is the semiaxis $\theta \geqslant 0$.

A statistical decision (statistical test) is performed by using some critical function $0 \leqslant \phi(x_1,x_2,\ldots,x_n) \leqslant 1$, defined on n-dimensional spaces (corresponding to the sample space). Specifically, the statistical decision is

(i) accept hypothesis K (i.e., reject hypothesis H) if $\phi(x_1,x_2,\ldots,x_n) = 1$;

(ii) accept hypothesis H (i.e., reject hypothesis K) if $\phi(x_1,x_2,\ldots,x_n) = 0$.

The cases where $0 < \phi(x_1,x_2,\ldots,x_n) < 1$ correspond to so-called randomized decisions. A randomized decision consists in performing a supplementary random experiment and accepting K with probability $\phi(x_1,x_2,\ldots,x_n)$ [i.e., rejecting H with probability $\phi(x_1,x_2,\ldots,x_n)$]. Nonrandomized tests only use values for $\phi(x_1,x_2,\ldots,x_n)$ equal to 0 or 1. Randomized tests are used [30] in situations when the critical function $\phi(x_1,x_2,\ldots,x_n)$ takes intermediate values between 0 and 1, which is often occurs for discrete random variables.

The power function $\beta_\phi(\theta)$ of the test ϕ is defined as the expectation of $\phi(x_1,x_2,\ldots,x_n)$ taken under the assumption that the parameter of the distribution generating the empirical sample is θ:

$$\beta_\phi(\theta) = \mathrm{E}_\theta[\phi(x_1,x_2,\ldots,x_n)] = \int\cdots\int \phi(x_1,x_2,\ldots,x_n) \times f(x_1|\theta)f(x_2|\theta),\ldots,f(x_n|\theta)dx_1,\ldots,dx_n. \quad (10)$$

By definition, the power function $\beta_\phi(\theta)$ is the probability to reject hypothesis H using the test $\phi(x_1,x_2,\ldots,x_n)$ with the parameter value of the distribution generating the empirical sample being θ.

The problem is to find tests ϕ that maximize the power $\beta_\phi(\theta)$ for all values of $\theta \in \Omega_K$, under the condition $\beta_\phi(\theta) \leqslant \gamma$ for all $\theta \in \Omega_H$, where γ is some small number that we choose preliminarily as an admissible false decision level for hypothesis H. Usually, γ is taken 0.10,0.05 or 0.01. Test $\phi(x_1,x_2,\ldots,x_n)$ is said more powerful than test $\psi(x_1,x_2,x_n)$ if

$$\beta_\phi(\theta) \geqslant \beta_\psi(\theta), \quad \text{for all} \quad \theta \in \Omega_K, \quad (11)$$

under the condition that both power functions take small enough values on Ω_H, i.e.,

$$\beta_\phi(\theta) \leqslant \gamma, \quad \text{for all} \quad \theta \in \Omega_H, \quad (12)$$

$$\beta_\psi(\theta) \leqslant \gamma, \quad \text{for all} \quad \theta \in \Omega_H. \quad (13)$$

The grail is to find the uniformly most powerful (UMP) test ϕ that is more powerful than any other test. Unfortunately, such test does not always exists. In particular, the UMP test does not exist for the problem of testing the Pareto (PL) distribution against the (truncated from below) lognormal (LN) distribution. But, if one adds some very reasonable restriction on the class of tests ϕ, then sometimes it is possible to find the UMP test in this narrower subclass. The restriction that sometimes ensures the existence of the UMP test is *unbiasedness*, which consists of the following. It is quite natural to add to the restriction (12) the following restriction:

$$\beta_\phi(\theta) \geqslant \gamma, \quad \text{for all} \quad \theta \in \Omega_K. \quad (14)$$

A test ϕ whose power function $\beta_\phi(\theta)$ satisfies conditions (12) and (14) is called an unbiased test. The condition (14) is natural, in the sense that, if it is not fulfilled, there will exist parameter values $\theta \in \Omega_K$ (i.e., under hypothesis K) for which the acceptance of hypothesis H is more likely than in some cases in which H is true.

It turns out that, in our problem of testing the Pareto vs LN distributions, it is possible to find the UMP test if one demands the fulfillment of the unbiasedness conditions (12) and (14). This test is nothing but the well-known maximum likelihood ratio test, with insertion in the maximum likelihood ratio of the maximum likelihood estimates of the unknown parameters instead of the true values. This test is known in the field of mathematical statistics as the Wilks test. It can be proven (see [26,29,30]) that the Wilks test for the problem of the Pareto vs LN distributions is a UMPU test. We can state that the most optimal statistical test in this problem (in the above-described sense) is the UMPU Wilks test. The test statistic equivalent to the Wilks test statistic can be chosen very simply: it is the sample coefficient of variation [the ratio of the sample standard deviation (std) to the sample mean]. The (minor) problem that is left is to determine the power function of the Wilks test. This problem can be solved either by the saddle point method, or by Monte Carlo simulations, as described below.

B. The uniformly most powerful unbiased test

As summarized in the introduction, one essential deficiency in statistical testing the LN vs PL hypotheses lies in the limited power of the used tests (L-test and χ^2-test, among others). While these tests are quite versatile, they are not

always very powerful. For instance, figure 2 in [14] which is reproduced as our Fig. 1 illustrates the lack of power of the L-test in the upper tail of the distribution under the null of a lognormal: the confidence bands derived from this test fan out very strongly, which makes this test unable to decide if the deviations observed in the data are genuine or fake. Of course, the main reason for the decreasing power observed in Fig. 1 (figure 2 in [14]) is the shrinking sample size for the upper ranks, but this does not remove the necessity of using the most possible powerful test in such a situation.

The discussion following Eqs. (4) and (5) suggests that, if the threshold u separating the LN from the PL is fixed, then it might be possible to clearly distinguish between the explanatory power offered by a lognormal distribution versus a Pareto distribution for the US Census 2000 data sample, when using a more powerful test for observations exceeding the threshold u. The most general and efficient test that addresses the core question, whether the Pareto law holds in the tail or the lognormal model is sufficient, is to consider for observations exceeding threshold u the two hypotheses:

H: Pareto distribution for values of x larger than some threshold u and

K: lognormal distribution also for value of x above the same threshold u.

Specifically, we propose to test the null hypothesis that, beyond some threshold u, the upper tail of the size distribution of cities or firms is Pareto

$$H : f_0(x;\alpha) = \alpha \cdot \frac{u^\alpha}{x^{\alpha+1}} \cdot 1_{x \geqslant u}, \quad \alpha > 0, \tag{15}$$

against the alternative that it is a (truncated from below) lognormal

$$K : f_1(x;\alpha,\gamma) = \frac{\gamma}{x} \frac{g(\alpha/\gamma)}{1 - G(\alpha/\gamma)} e^{-\alpha \ln(\frac{x}{u}) - \frac{\gamma^2}{2}[\ln(\frac{x}{u})]^2} \cdot \mathbf{1}_{x \geqslant u}, \tag{16}$$

where $\alpha \in \mathbb{R}, \gamma > 0$, and the functions $g(\ldots)$ and $G(\ldots)$ are the standard normal probability density function (PDF) and cumulative distribution function (CDF), respectively. Note that the threshold u does not have to be equal to some data point and can take any continuous value.

This is equivalent to testing the null hypothesis that the upper tail of the log-size distribution of cities or firms is exponential against the alternative that it is a (truncated) normal. For this latter problem, Ref. [26] has shown that the clipped sample coefficient of variation $\hat{c} = \min(1,c)$, which is equivalent for this problem to the Wilks likelihood ratio statistic, provides the uniformly most powerful unbiased test. Here, c is the sample coefficient of variation of $\ln(x/u)$ defined as the ratio of the sample standard deviation to the sample mean of the random variable $\ln(x/u)$. The standard deviation is the square root of the variance, and the variance is estimated using the usual Bessel correction giving an unbiased estimator of the population variance. Note that the use of the likelihood ratio Wilks test corresponds to a "nested" test, i.e., the power-law can be seen as a limit case of the lognormal distribution.

With the notations of the general theory of statistical hypotheses testing used in Sec. III A, the clipped sample coefficient of variation plays the role of a sufficient statistic for the optimal unbiased UMP test. The critical function $\phi(x_1,x_2,\ldots,x_n)$ reads

$$\phi(x_1,x_2,\ldots,x_n) = 1 \text{ if } \hat{c} < h_\gamma; \text{ we reject PL and accept LN;}$$
$$\phi(x_1,x_2,\ldots,x_n) = 0 \text{ if } \hat{c} \geqslant h_\gamma; \text{ we reject LN and accept PL.}$$

The threshold h_γ is chosen such that the probability that the inequality $\hat{c} < h_\gamma$ holds under the null hypothesis (PL) is equal to the (small) preliminarily chosen value γ, typically chosen equal to 0.1, 0.05, or 0.01.

The critical threshold h_γ of the test can be derived with extremely high accuracy (even for very small samples) by a saddle point approximation [26,27] or by Monte Carlo methods. The Monte Carlo estimation is very simple and proceeds as follows. First, it is necessary to generate M samples distributed according to the standard exponential law (with unit parameter), i.e., $\exp(-x)$, and of a fixed size equal to the number of observations exceeding the given threshold u of the real sample. The number M should be taken large enough, say, $M = 10\,000$. Then, for each sample, the clipped coefficient of variation (CCV) is calculated and compared with that of the real sample. The fraction of exceedances provides a good statistical estimate of the corresponding p-value of the null hypothesis (PL).

In our problem, the p-value is the probability to reject the null PL hypothesis when it is true. In the general notations of Sec. III A, this probability was denoted as $\beta_\phi(\theta)$. The smaller the p-value corresponding to the observed value of the test statistic, the more likely is the PL hypothesis and the less likely is the LN hypothesis. Usually, p-values equal to 0.10 or smaller can be considered small enough for practical problems, and sufficient to reject the alternative hypothesis.

From Fig. 1, one can see a deviation between the empirical complementary distribution function and the LN fit, which becomes important at about $\exp(10.5) \simeq 36\,000$ inhabitants. This suggests that the appropriate threshold u should be somewhere near this value. In order to determine the most appropriate value of the threshold u separating the LN from the PL distributions, we suggest to use the maximum likelihood estimation, which as seen below confirms the visual inspection.

The likelihood L for the whole sample is

$$L(x_1,x_2,\ldots,x_n|\alpha,\gamma,u,\alpha_0) = p^{n_1} \cdot L_1 \cdot (1-p)^{n_2} \cdot L_2, \tag{17}$$

where α,γ are LN parameters for the lower part of the sample; u is the threshold separating the LN from the PL distributions; p is the probability of not exceeding threshold u; n_1 and n_2 are the numbers of observations, respectively, below and above u (with $n_1 + n_2 = n$); α_0 is the PL parameter for the upper part of the sample; L_1 and L_2 are, respectively, the likelihoods for the lower part and the upper part of the sample:

$$L_1(x_1,x_2,\ldots,x_n; x_k \leqslant u, 1 \leqslant k \leqslant n|\alpha,\gamma,u) = \prod_{x_i \leqslant u} f(x_i|\alpha,\gamma,u); \tag{18}$$

$$f(x|\alpha,\gamma,u) = \frac{\gamma}{x} \frac{g(\alpha/\gamma)}{G(\alpha/\gamma)} e^{-\alpha \ln(\frac{x}{u}) - \frac{\gamma^2}{2}\left[\ln(\frac{x}{u})\right]^2}; x \leqslant u. \tag{19}$$

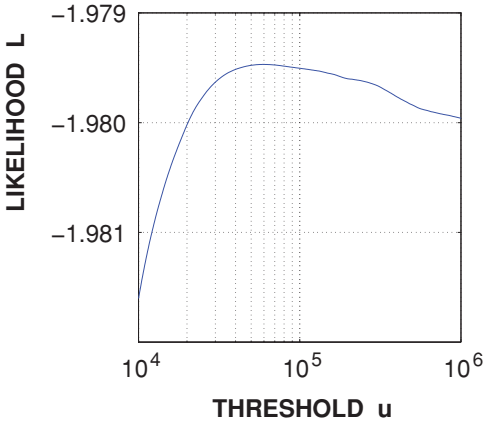

FIG. 3. (Color online) Likelihood $L(x_1,x_2,\ldots,x_n|\alpha^*,\beta^*,u,\alpha_0^*)$ defined by (17) and following equations as a function of the threshold u (decimal logarithmic scale) where $(\alpha^*,\beta^*,\alpha_0^*)$ are the MLE of the parameters (α,β,α). The maximum occurs at $u^* = 37\,235$ inhabitants, providing the most appropriate estimate of the threshold u separating the LN from the PL regimes.

The functions $g(\ldots)$ and $G(\ldots)$ are the standard normal probability density function (PDF) and cumulative distribution function (CDF), respectively:

$$L_2(x_1,x_2,\ldots,x_n; x_k > u, 1 \leqslant k \leqslant n|u,\alpha_0) = \prod_{x_i>u} f_0(x_i|u,\alpha_0); \tag{20}$$

$$f_0(x_i|u,\alpha_0) = \alpha_0 \frac{u^{\alpha_0}}{x^{1+\alpha_0}}, x > u. \tag{21}$$

The likelihood $L[x_1,x_2,\ldots,x_n|\alpha^*(u),\gamma^*(u),u,\alpha_0^*(u)]$ is shown in Fig. 3 as a function of the threshold u where $[\alpha^*(u),\gamma^*(u),\alpha_0^*(u)]$ are the MLE of the parameters (α,γ,α_0), which are functions of the threshold u. One can observe that L reaches its maximum at $\ln u^* = 11.0$ (i.e., $u^* = 37\,235$ inhabitants). The composite likelihood L has a plateau near its maximum, so that thresholds within the interval $10.5 \leqslant \ln u \leqslant 11.5$ have practically the same likelihood value, and one can choose the lowest value $\ln u^* = 10.5$ ($u^* = 36\,316$) as providing the maximum number of observations for the upper part of the sample (the Paretian domain), therefore extending the domain where there is a relative scarcity of observations for the estimation of parameter α_0. For threshold $u^* = 36\,316$, we obtain the following ML estimates of the parameters: $\alpha^* = 1.094; \gamma^* = 0.581; \alpha_0^* = 1.161$. There are 977 observations exceeding this threshold $u^* = 36\,316$.

C. Distinguishing Pareto tail and lognormal distribution

1. *The distribution of city sizes in the US Census 2000 data*

The p-value for testing the LN vs PL distributions for observations exceeding this threshold $u^* = 36\,316$ with 977 observations, as determined in the previous subsection by maximum likelihood, is equal to 5% and remains smaller than 5% for larger population thresholds. Thus, we can state that, for the tail of city size distribution (with 977 largest cities, containing 56.85% of the whole population summed over all cities), the hypothesis LN should be rejected, and the hypothesis PL should be accepted.

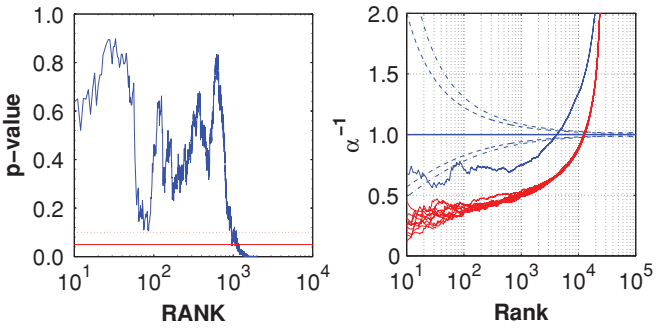

FIG. 4. (Color online) The left panel depicts the p-value of the test of the null hypothesis that the upper tail of the size distribution of cities is Pareto against the alternative that it is a (truncated) lognormal as a function of the rank (decimal logarithmic scale), where cities are ordered by decreasing sizes. The right panel depicts Hill's estimate of the inverse of the tail index for the Census 2000 data (blue upper curve) and for ten samples drawn from a lognormal distribution with parameters $\mu = 7.28$ and $\sigma = 1.25$ (red bottom curves) as a function of city ranks (decimal logarithmic scale). The two dashed (respectively, dot-dash) curves provides the confidence bands at the 5%-significance level (respectively, 1% level) derived from the UMPU test that the tail index $\alpha = 1$ against a two-sided alternative.

The left panel of Fig. 4 depicts the p-value for the reverse test of the hypothesis H (PL) versus hypothesis K (LN), as a function of the lower threshold u expressed in terms of the rank of city sizes of the US census 2000 represented in a logarithmic scale. The p-values have been calculated using the saddle-point approximation [26,27]. Extensive Monte Carlo simulations, performed as explained in the previous section, reproduce basically the same results. Figure 4 indubitably confirms again that one cannot reject the hypothesis that the size distribution of the 1000 largest cities or so, which include more than half of the total population, is Pareto. This confirms and makes more precise the claim in [10]. For larger ranks (smaller thresholds), the p-value becomes very small, which leads to the rejection of the Pareto distribution, and the need for the lognormal distribution to describe the set of smaller cities. This explains Eeckhout's results [13].

The right panel of Fig. 4 depicts Hill's estimate $\hat{\alpha^{-1}}$ of the inverse of the tail index α of the Pareto distribution (15) again as a function of city rank. This estimator is the best unbiased estimator for the inverse of the tail index[4] [28]. For the US census 2000 data (blue upper noisy curve), the inverse of the tail index is approximately constant and fluctuates around the value 0.7 for ranks less than one thousand or so, confirming the validity of the Pareto model over this range. For ranks larger than one thousand, the Hill's estimate $\hat{\alpha^{-1}}$ deviates rapidly, confirming a deviation from the Pareto model for the set of smaller cities.

In the right panel of Fig. 4, we also show Hill's estimate $\hat{\alpha^{-1}}$ for ten random samples drawn from a lognormal distribution with parameters $\mu = 7.28$ and $\sigma = 1.25$ (red curves). One can observe the absence of a plateau, and therefore no well-defined

[4]It is not possible to get an unbiased estimate for α.

exponent, thus disqualifying the Pareto model to approximate data generated by the pure lognormal distribution estimated in [13]. In contrast, the real curve obtained with the empirical data exhibits a definite plateau in the range of ranks 10–10^3, corresponding approximately to a Pareto index $\alpha \approx 1.4$. The increase of $\hat{\alpha^{-1}}$ with rank is the expected signature of the fact that the lognormal density is rapidly decreasing, i.e., it goes to zero faster than any power-law, so that its effective tail index tends to infinity and its inverse is vanishing. Therefore, for the highest ranks (largest cities), Hill's estimator should converge to zero for data generated by a lognormal distribution, just as we see on ten curves shown in Fig. 4 corresponding to lognormal samples.

The contrast between the US Census 2000 data and the samples drawn from a lognormal distribution with parameters $\mu = 7.28$ and $\sigma = 1.25$ is striking and provides additional evidence in favor of the Pareto distribution for the upper tail. This makes clear that the Pareto and lognormal models are distinguishable in their tail for the available US Census 2000 data sample.

2. The distribution of identity losses

In order to demonstrate that our procedure works for data sets other than the US city size distribution extensively studied here, we briefly present the results obtained for a completely different data set, previously analyzed in Ref. [31], which exhibits a power-law tail together with a lognormal-like shape in the bulk of the distribution.

The data set consists in a catalog of personal identity (ID) losses. ID loss event data have been thoroughly collected by several independent organizations. As in Ref. [31], we use the most complete data set from the Open Security Foundation [http://datalossdb.org/ (06.01.2009)], that contains 956 documented events reported mainly in the USA between year 2000 and November 2008. An event is defined following the procedure described in Ref. [32]. For instance, the largest entries in the data set are (i) the discovery and disclosure of an attack over several years of the TJX Companies[5] with a probable exposition of more than 9×10^7 IDs (end of the event: January 2007), (ii) the Cardsystems' hack impacting 4×10^7 Visa, MasterCard and American Express cardholders (June 2005), (iii) America Online (3×10^7 credit card ID exposed in 2004), and (iv) the US Department of Veterans Affair (more than 2.5×10^7 of ID stolen in 2006). The catalog provides also the involved organization, the date and amount of loss (measured as the numbers of ID stolen). Data are homogeneously sampled among various types of organizations: business (35%), education (30%), governments (24%), and medical institutions (10%).

Figure 5 shows the empirical complementary cumulative distribution function (ccdf) of the ID losses over this catalog of 956 events. For events with ID losses smaller than the threshold $u^* = 299\,540$ indicated by the vertical segment, the empirical ccdf seems to be well fitted by a lognormal law shown as the continuous thin line curving downward. For events with ID losses larger than $u^* = 299\,540$, the empirical ccdf seems to depart significantly from the lognormal law and is well fitted by a power law tail shown as the straight line in this log-log plot.

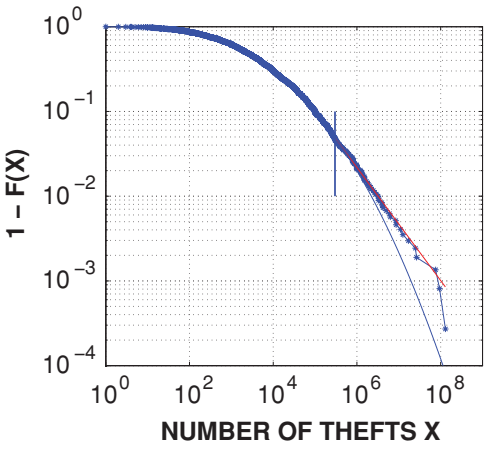

FIG. 5. (Color online) Empirical complementary cumulative distribution function (ccdf) (decimal logarithmic scale) of the ID losses (decimal logarithmic scale) over the catalog of 956 events used previously in Ref. [31]. The vertical segment indicates the threshold $u^* = 299\,540$ above which the ccdf is qualified as a power-law.

The threshold $u^* = 299\,540$ mentioned above has been obtained by maximum likelihood estimation using the form of the likelihood function given by expression (17) and the procedure described in Sec. III B. As in Fig. 3 for the city size distribution, the likelihood $L(x_1,x_2,\ldots,x_n|\alpha^*,\gamma^*,u,\alpha_0^*)$ of the set of ID losses is shown in Fig. 6 as a function of the threshold u where $(\alpha^*,\gamma^*,\alpha_0^*)$ are the MLE of the parameters (α,γ,α_0). The set of ID loss events with losses larger than $u^* = 299\,540$ contains just 90 events. The ML estimates of the parameters defining the lognormal regime below u^* and the power-law regime above u^* are: $\alpha^* = 0.625; \gamma^* = 0.350; \alpha_0^* = 0.667$. The later value confirms with better precision the previous estimate $\alpha_0^* = 0.7 \pm 0.1$ of Ref. [31]. Here, in addition, we can state that the p-value for the null hypothesis of the lognormal model is so small for the largest ranks (the p-value for ranks larger than 300 reaches 5% and remains smaller than 0.01% for ranks larger than 450) so as to lead to the conclusion that the lognormal hypothesis should be rejected, and the power-law hypothesis should be accepted.

Finally, Fig. 7 is the same as the left panel of Fig. 4 and shows the p-value for the test of the hypothesis H (PL) versus hypothesis K (LN), as a function of the lower threshold u expressed in terms of the rank of the ID loss size represented in a logarithmic scale. The p-values have been calculated using the saddle point approximation [26,27], and confirmed by extensive Monte Carlo simulations. Figure 7 shows that one cannot reject the hypothesis that the size distribution of the 300 largest ID losses, is Pareto. For larger ranks (smaller thresholds), the p-value becomes very small, which leads to the rejection of the Pareto distribution, and the need for the lognormal distribution to describe the set of smaller ID losses. This p-value calculation is a bit more optimistic in its assessment of the range where the PL hypothesis holds,

[5]The TJX Companies, Inc. is a large retailer of apparel and home fashions in the United States and worldwide.

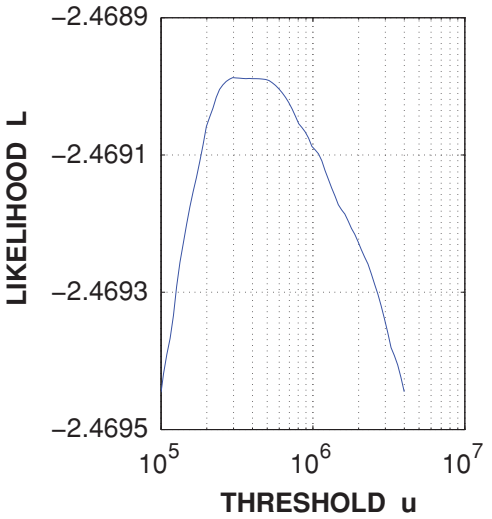

FIG. 6. (Color online) Same as Fig. 3 for the cyber-risk data set consisting in a catalog of identity loss events. The maximum occurs at $u^* = 299\,540$ ID losses, providing the most appropriate estimate of the threshold u separating the log normal (LN) from the power-law (PL) regimes. The thresholds u in the abscissa are represented with a decimal logarithmic scale.

compared with the ML estimate which gives a shorter PL tail (rank 1 to about rank 90).

D. Pareto model versus Zipf's law in the US Census 2000 data

Now that we have established that the tail of the size distribution of cities is compatible with the Pareto distribution (which should be selected as the most parsimonious hypothesis), we turn to the question of whether this Pareto law is Zipf's law, i.e., whether the exponent is $\alpha = 1$.

First, the right panel of Fig. 4 shows the confidence band at the 95% and 99% significance levels for the hypothesis that the tail exponent of the PL is equal to 1 (Zipf's law). In other words, the upper and lower 95% confidence limits (correspondingly, 99% limits) delineate the domain within which the true value of the parameter α^{-1} can occur with probability 95% (correspondingly, 99%) under the assumption

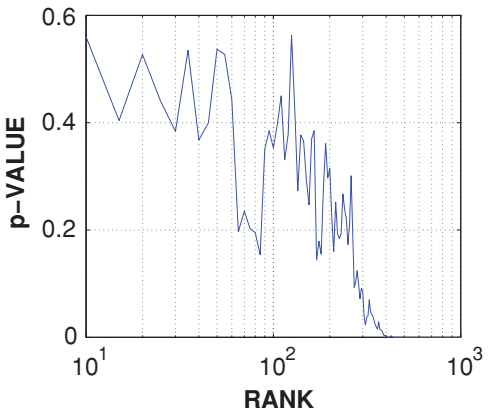

FIG. 7. (Color online) Same as left panel of Fig. 4 for the catalog of ID losses.

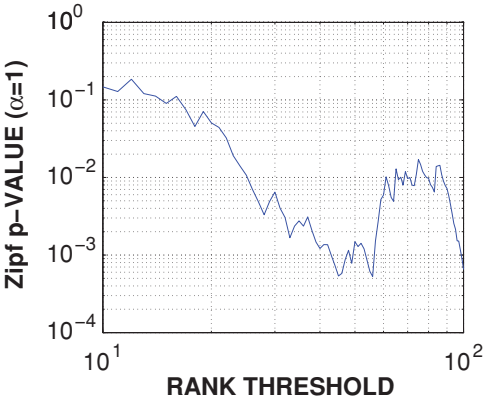

FIG. 8. (Color online) One-sided p-value (decimal logarithmic scale) as a function of rank threshold (decimal logarithmic scale), testing the hypothesis that the tail exponent of the Pareto distribution is compatible with Zipf's laws that $\alpha = 1$. The p-value is defined as the probability of exceeding the observed index estimate under the hypothesis that Zipf's law holds, i.e., that $\alpha = 1$.

that the city size distribution follows Zipf's law, i.e., $\alpha = 1$. At the 95% significance level, Zipf's law is rejected, except for the 20 largest cities.

Figure 8 improves on this statistics by plotting the p-value defined as the probability of *exceeding* the observed index estimate (one side-test) under the hypothesis that Zipf's law holds (index equals to unity). For rank thresholds larger than 20, all p-values are smaller than 0.05. For rank thresholds larger than 16, all p-values are smaller than 0.10. We are thus led to conclude that Zipf's law cannot be accepted to describe the tail of the distribution of city sizes in the US census studied here, whereas a larger exponent approximately equal to 1.4 is significantly more likely.

IV. CONCLUDING REMARKS

We have proposed the uniformly most powerful unbiased (UMPU) test between the lognormal and the power-laws, as a general statistical tool to use systematically when researchers encounter a fat-tail probability distribution function. Because the power-law model is often argued to be a general property exhibited by natural and social complex systems, and because the lognormal distribution is also ubiquitously associated with proportional growth processes with fat-tail properties which make it difficult to distinguish from the power-law especially in standard log-log plots, the UMPU test should become in our opinion the standard tool for assessing the nature of the tail of empirical distributions.

We have presented a pedagogical introduction which motivates and shows how the UMPU test between the lognormal and the power-laws is constructed. We have introduced a maximum likelihood estimation (MLE) of the threshold u separating a possible lognormal regime in the bulk of the distribution from a putative power-law regime in the tail. This has allowed us to test directly the power-law versus lognormal hypothesis in a predefined sample given by the data sizes larger than the threshold.

We have presented two applications on empirical data sets. The first one consists in the distribution of US city sizes, for which there is an unsettled controversy between Eeckhout [13,14] and Levy [10], concerning the validity of Zipf's law. By using the UMPU test between the lognormal and the power-laws, we have shown that conclusive results can be achieved to end this debate. We can state that, for the tail of city size distribution (with 977 largest cities, containing 56.85% of the whole population summed over all cities), the lognormal hypothesis should be rejected, and the power-law hypothesis should be accepted. However, we exclude the Zipf exponent $\alpha = 1$ and find that the Pareto exponent is equal to 1.4 ± 0.1. A review of the mechanisms based on Gibrat's law leading to distributions with Pareto tails whose exponents can deviate from the Zipf's law value $\alpha = 1$ can be found in [22,33].

In order to demonstrate that our procedure works for data sets other than the US city size distribution, we have also presented a brief analysis of the power-law tail of the distribution of personal identity (ID) losses, which constitute one of the major emergent risks at the interface between cyberspace and reality, which was previously analyzed in Ref. [31].

ACKNOWLEDGMENTS

We acknowledge financial support from the ETH Competence Center "Coping with Crises in Complex Socio-Economic Systems" (CCSS) through ETH Research Grant CH1-01-08-2 and ETH Zurich Foundation.

APPENDIX

Let us consider a sample of i.i.d. (independently identically distributed) random variables with continuous distribution function (DF) $F(y)$ and sample size n. Let us denote the empirical DF as $F_n(y)$. Then, the Kolmogorov distance D_n, defined as

$$D_n = n^{1/2} \cdot \max|F_n(y) - F(y)|, \tag{A1}$$

has asymptotically the Kolmogorov distribution, independently of $F(y)$. Assuming the validity of the Kolmogorov distribution for D_n, i.e., that n is large enough, the quantile q_p defined by

$$P\{D_n < q_p\} = p, \tag{A2}$$

can be determined easily for any desired p, e.g., $p = 0.95$. The inequality $D_n < q_p$ is equivalent to the following chain of two inequalities:

$$1-F(y)-\frac{q_p}{n^{1/2}} < 1-F_n(y) < 1-F(y)+\frac{q_p}{n^{1/2}}, \quad \text{for all } y. \tag{A3}$$

Using (A2) and (A3), we get

$$P\left\{1 - F(y) - \frac{q_p}{n^{1/2}} < 1 - F_n(y) < 1 - F(y) + \frac{q_p}{n^{1/2}}\right\} = p. \tag{A4}$$

We can interpret (A4) as a confidence interval at the confidence level p for the empirical tail $1 - F_n(y)$. Taking the logarithm of the right and left sides of the inequalities in (A4), we have

$$P\left\{\ln\left(1 - F(y) - \frac{q_p}{n^{1/2}}\right) \right.$$
$$\left. \langle\ln[1 - F_n(y)]\langle\ln\left(1 - F(y) + \frac{q_p}{n^{1/2}}\right)\right\} = p. \tag{A5}$$

It is clear from (A5) that, when the tail $1 - F(y)$ becomes equal to or less than $\frac{q_p}{n^{1/2}}$, the true log-tail $\ln[1 - F(y)]$ deviates from $\ln[1 - F(y) \pm \frac{q_p}{n^{1/2}}]$ very strongly, since the term $1 - F(y)$ becomes less than the term $\frac{q_p}{n^{1/2}}$ leading to a divergence of the logarithm. Hence, we can say that the very tail $1 - F(y)$ cannot be distinguished well on the "background" of the constant term $\frac{q_p}{n^{1/2}}$. This "eclipse" occurs for y satisfying the following condition:

$$1 - F(y) = c\frac{q_p}{n^{1/2}}, \tag{A6}$$

where c is some constant, e.g., $c \simeq 2$.

The Lilliefors test modifies the Kolmogorov test in the situation when the theoretical DF $F(y)$ is the Gaussian distribution function $G(y|m,s)$ with unknown parameters (m,s) that are replaced in the equations used above by their sample estimates [sample mean and sample standard deviations (std)]. In accordance with above remark, the efficiency of the Lilliefors test becomes very low as $1 - G(y|m,s)$ becomes smaller than $2\frac{q_p}{n^{1/2}}$. In other words, the Lilliefors test "works" efficiently in the middle range, whereas it fails completely in the tail range.

Applied to the city size distribution, we have $n = 25356, Y = \ln(X)$, where X is a city size, $m = \text{mean}(Y) = 7.2781; s = \text{std}(Y) = 1.7529$. Using $p = 0.95$, the theoretical 0.95 quantile for $\max |F_n(y)\eta G(y|m,s)|$ is equal to 0.0057. The observed value is $\max |F_n(y) - G(y|m,s)| = 0.0190$. Under the null H_0 hypothesis (lognormal distribution of city sizes), the probability that $\max |F_n(y) - G(y|m,s)|$ exceeds 0.0190 is less than 0.001. Thus, in accordance with the Lilliefors test, we should reject H_0. Thus, Eeckhout was not correct when he used the Lilliefors test in order to support H_0 in the tail range and did not say that this test rejects the hypothesis H_0 on the whole range. Indeed, detailed observation of Fig. 1 shows that the 95% confidence interval covers the empirical DF $F_n(y)$ in the tail range, but there are some other places along the distribution for which the empirical DF $F_n(y)$ goes out of the 95%-confidence interval.

[1] D. Sornette, in *Probability Distributions in Complex Systems*, edited by R. A. Meyers (Springer Science, Berlin, 2009), Vol. 11, p. 4300.

[2] P. Bak, *How Nature Works: the Science of Self-organized Criticality* (Copernicus, New York, 1996).

[3] D. Sornette, *Critical Phenomena in Natural Sciences, Chaos, Fractals, Self-organization and Disorder: Concepts and Tools*, 2nd ed., Springer Series in Synergetics, (Springer, Heidelberg, 2006).

[4] D. Sornette, J. Phys. I (France) **4**, 209 (1994).

[5] D. Sornette, Int. J. Mod. Phys. C **13**, 133 (2002).
[6] M. E. J. Newman, Contemp. Phys. **46**, 323 (2005).
[7] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley Press, Cambridge, MA, 1949).
[8] R. L. Axtell, Science **293**, 1818 (2001).
[9] X. Gabaix, Quarterly Journal of Economics **114**, 739 (1999).
[10] M. Levy, American Economic Review **99**, 1672 (2009).
[11] E. G. J. Luttmer, Quarterly Journal of Economics **122**, 1103 (2007).
[12] L. M. B. Cabral and J. Mata, American Economic Review **93**, 1075.
[13] J. Eeckhout, American Economic Review **94**, 1429 (2004).
[14] J. Eeckhout, American Economic Review **99**, 1676 (2009).
[15] R. Gibrat, *Les Inégalités Economiques; Applications aux inégalitiés des richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel* (Librarie du Recueil Sirey, Paris, 1931).
[16] E. Rossi-Hansberg and M. L. J. Wright, Urban Structure and Growth, Review of Economic Studies **74**, 597 (2007).
[17] Z. Chen, S. Fu, and D. Zhang, Searching for the Parallel Growth of Cities (SSRN series [http://papers.ssrn.com/sol3/papers.cfm?abstract]).
[18] P.-P. Combes, G. Duranton, L. Gobillon, D. Puga, and S. Roux, The Productivity Advantages of Large Cities: Distinguishing Agglomeration from Firm Selection (SSRN series [http://papers.ssrn.com/sol3/papers.cfm?abstract_id=135640]).
[19] H. W. Lilliefors, J. Am. Stat. Assoc. **62**, 399 (1967).
[20] M. A. Stephens, J. Am. Stat. Assoc. **69**, 730 (1974).
[21] N. H. Bingham, C. M. Goldie, and J. L. Teugels, *Regular Variation* (Cambridge University Press, Cambridge, 1989).
[22] A. Saichev, Y. Malevergne, and D. Sornette *Theory of Zipf's Law and Beyond*, Lecture Notes in Economics and Mathematical Systems 632 (Springer Saichev, New York, 2010).
[23] A. N. Kolmogorov, Doklady AN USSR **31**, 99 (1941) (in Russian).
[24] D. Sornette, Physica A **250**, 295 (1998).
[25] H. Kesten, Acta Mathematica **131**, 207 (1973).
[26] J. Del Castillo and P. Puig, J. Am. Stat. Assoc. **94**, 529 (1999).
[27] R. Gattoa and S. R. Jammalamadakab, Statistics and Probability Letters **58**, 71 (2002).
[28] B. M. Hill, Annals of Statistics **3**, 1163(1975).
[29] E. L. Lehmann, *Testing Statistical Hypotheses* (Chapman and Hall, New York, 1994).
[30] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 3rd ed. (Springer, New York, 2005).
[31] T. Maillart and D. Sornette, Eur. Phys. J. B **75**, 357 (2010).
[32] R. Hasan and W. Yurcik, "A statistical analysis of disclosed storage security breaches," *Proceedings of the second ACM workshop on Storage security and survivability* (2006), pp. 1–8.
[33] Y. Malevergne, A. Saichev, and D. Sornette, [http://ssrn.com/abstract=1083962].