Pareto versus lognormal: A maximum entropy test

Marco Bee*
Department of Economics, University of Trento, Trento, Italy

Massimo Riccaboni†
DISA, University of Trento, Trento, Italy and IMT School of Advanced Studies, Lucca, Italy

Stefano Schiavo‡
Department of Economics, University of Trento, Trento, Italy and OFCE-DRIC, Valbonne, France

(Received 11 February 2011; revised manuscript received 23 May 2011; published 4 August 2011)

It is commonly found that distributions that seem to be lognormal over a broad range change to a power-law (Pareto) distribution for the last few percentiles. The distributions of many physical, natural, and social events (earthquake size, species abundance, income and wealth, as well as file, city, and firm sizes) display this structure. We present a test for the occurrence of power-law tails in statistical distributions based on maximum entropy. This methodology allows one to identify the true data-generating processes even in the case when it is neither lognormal nor Pareto. The maximum entropy approach is then compared with other widely used methods and applied to different levels of aggregation of complex systems. Our results provide support for the theory that distributions with lognormal body and Pareto tail can be generated as mixtures of lognormally distributed units.

DOI: 10.1103/PhysRevE.84.026104 PACS number(s): 89.65.Gh, 02.50.Ng, 05.40.−a

I. INTRODUCTION

Several phenomena in physics, biology, computer science, demography, economics, finance, and the social sciences are distributed according to a power law, or at least display power-law behavior in the tails [1–11]. The power-law upper tail of the distribution can be generated by an amplification method [2], such as mixtures of lognormals [5,12,13]. In the last decade the debate has intensified on the appropriate procedures to detect power-law distributions in empirical data [14–17], and a number of approaches have been proposed to establish the length of the power-law tail [18,19], quickly gaining widespread acceptance and use. In the literature the power-law (Pareto) distribution is generally compared to an alternative represented by the lognormal, though other candidate distributions have been proposed [6,17,20–22]. While in many cases the exact shape of the empirical distribution is not crucial, as long as heavy tails are accounted for, the debate appears to be especially animated in physics [7,9,11,23–28], economics [29–35], and biology [12,36]. Further complications come from the fact that there is no unique definition of heavy-tailed distributions [37], and many social and natural phenomena may display different tail behaviors when analyzed at different levels of aggregation, due to composition and sample size effects [12,38–40].

In this paper we provide a methodology based on maximum entropy (ME) estimation [41,42] to identify the data-generating process and to determine the existence of a power-law tail in the data.[1] Two of the main benefits of this approach are its flexibility and the fact that it delivers a well-defined alternative to the power-law or lognormal distribution. As the ME density encompasses most commonly used distributions, the estimated ME density can be easily compared with a number of alternatives. Here we apply the ME methodology to evaluate the fit of lognormal versus power-law distributions, to compare different systems, and to analyze the behavior of the same complex systems at different levels of aggregation.

In what follows we briefly describe the theoretical framework associated with heavy-tailed distributions, review the most commonly used methodologies to estimate the upper-tail behavior of empirical data, and introduce the ME approach. Then we analyze the distributions of city size, world trade flows, and business firm size. In the last case we find support for a theoretical prior suggesting the emergence of a power-law upper tail in the distribution upon aggregation [13,38,43]. Finally, we compare the results of different tests by means of simulations.

II. ESTIMATING HEAVY-TAILED DISTRIBUTIONS

There are several definitions of heavy-tailed distributions [37]. In applications, the two most commonly used heavy-tailed distributions are the lognormal and the power-law. Both of them are heavy-tailed according to the following definition [37, p. 50], [44, p. 5]:

Definition 1. A random variable X with cumulative distribution function (CDF) F on $(0,+\infty)$ is heavy-tailed if $\mathrm{E}(e^{tX}) = \int_0^\infty e^{tx}\, dF(x) = \infty, \forall t \in \mathbb{R}$.

Conversely, the lognormal does not belong to the class identified by the most restrictive definition [37, p. 564], which identifies a power-law tail:

Definition 2. A random variable X with CDF F on $(0,+\infty)$ is heavy-tailed if there exists a positive parameter α such that $\lim_{x\to\infty} \bar{F}(x)/x^{\alpha} = L(x)$, where $\bar{F}(x) = 1 - F(x)$ and $L(x)$ is a slowly varying function.

Distributions that satisfy Definition 2 are a subset of distributions satisfying Definition 1.

*marco.bee@unitn.it
†massimo.riccaboni@unitn.it
‡stefano.schiavo@unitn.it

[1]The MATLAB code implementing the methodology is available at [www.stefanoschiavo.tk].

The tail behavior of the lognormal and Pareto distributions can be explained by extreme value theory (EVT). Given a random variable with CDF F and some predefined large value $x_{\min} \in \mathbb{R}^+$ in the support of F, define the excess distribution over the threshold $x_{\min}$ as $Y = X - x_{\min}$. The probability that $X > x_{\min}$ by no more than an amount $y \geqslant 0$, given that the threshold has been exceeded, is

$$P(X - x_{\min} \leqslant y | X > x_{\min}) = F_{x_{\min}}(y) = \frac{F(y + x_{\min}) - F(x_{\min})}{1 - F(x_{\min})}, \quad 0 \leqslant y < x_0 - x_{\min},$$

where $x_0 \leqslant \infty$ is the right endpoint of F. The Balkema, de Haan, and Pickands (BHP) theorem guarantees that under some conditions there is a function $\beta(x_{\min})$ such that the excess distribution converges to the generalized Pareto distribution (GPD) [37]:

$$\exists\, \beta(x_{\min}) > 0 : \lim_{x_{\min} \to x_0} \sup_{0 \leqslant y < x_0 - x_{\min}} |F_{x_{\min}}(y) - G_{\xi,\beta(x_{\min})}(y)| = 0 \iff F \in \mathrm{MDA}(H_\xi),\ \xi \in \mathbb{R},$$

where

$$G_{\xi,\beta}(y) = \begin{cases} 1 - \left(1 + \xi \frac{y}{\beta}\right)^{-\frac{1}{\xi}} & \text{if} \quad \xi \neq 0, \\ 1 - \exp\left(-\frac{y}{\beta}\right) & \text{if} \quad \xi = 0, \end{cases}$$

$$S(G_{\xi,\beta}) = \begin{cases} y \geqslant 0 & \text{if} \quad \xi \geqslant 0, \\ 0 \leqslant y \leqslant \frac{\beta}{\xi} & \text{if} \quad \xi < 0, \end{cases}$$

are, respectively, the CDF of the GPD and its support, and $\mathrm{MDA}(H_\xi)$ is the maximum domain of attraction of the generalized extreme value distribution with shape parameter ξ. This means that the distribution of $Y = X - x_{\min}$ is well approximated by a GPD, for sufficiently large $x_{\min}$, and that the GPD parameter β is a function of $x_{\min}$. The lognormal distribution converges to $G_{\xi,\beta}$ with $\xi = 0$ (Gumbel-type), whereas the Pareto [denoted $\mathrm{Par}(c,\alpha)$, where c is the scale and α the shape parameter] to $G_{\xi,\beta}$ with $\xi = 1/\alpha > 0$ (Fréchet-type). This implies that the asymptotic tail behaviors of the two distributions are different. However, the convergence is very slow, so that, as shown with different arguments in Refs. [39] and [19], the difference may be very small, at the extent that they are often practically indistinguishable for any finite sample size.

The classical approach to the estimation of the parameters of the Pareto distribution is based on a random sample from the $\mathrm{Par}(c,\alpha)$ distribution. The maximum likelihood estimators (MLEs) of the parameters are [45]

$$\hat{c} = \min_{1 \leqslant i \leqslant n} x_i; \quad \hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \log(x_i/\hat{c})}. \tag{1}$$

However, since we know that only observations larger than some unknown threshold $x_{\min}$ follow the Pareto distribution, the threshold cannot be estimated by means of Eq. (1). In this case, a two-step procedure is often applied [37]: (1) plot the mean excess function $m = E(X - x | X > x)$; and (2) letting $\Delta_n(x) = \{i : X_i > x\}$, set $x_{\min} = x^*$, where x^* is the smallest number such that the empirical mean excess function $\hat{m} = [1/\#\Delta_n(x)] \sum_{i \in \Delta_n(x)} (X_i - x)$ (where # denotes the cardinality of a set) is approximately linear for $x > x^*$.

Alternatively, one may use the Hill estimator [37,46], which is equivalent to the MLE of the shape parameter α if the underlying distribution is Pareto. When the tail is Pareto above a certain threshold $x_{\min}$, the Hill estimator is the MLE conditional on the threshold being equal to the kth-order statistic $x_{(k)}$. By plotting the Hill estimator as a function of $x_{(k)}$ (Hill plot) one can determine $x_{\min}$ so that the plot is approximately linear for $x_{(k)} \geqslant x_{\min}$, since the Hill estimator is the empirical mean excess function of $\log(x)$ computed at $x_{\min} = \log(x_{(k)})$ [37]. However, the Hill estimator in finite samples can be severely biased [15,37].

Another approach is based on the fact that the logarithm of a (truncated) lognormal is a (truncated) normal, and the logarithm of a Pareto is an exponential. The likelihood ratio test for the null hypothesis of exponentiality against the alternative of truncated normality is given by the clipped sample coefficient of variation $\bar{c} = \min\{1, \hat{\sigma}/\hat{\mu}\}$ (where μ and σ are the parameters of the truncated normal). It is also the uniformly most powerful unbiased (UMPU) test [19,40,47]. This test is computationally simple and theoretically appealing, but the restriction of the UMP property to the class of unbiased tests is often not completely satisfactory from a statistical point of view [48]. Moreover, the UMPU test can be used for testing only the Pareto versus the lognormal distribution.

Recently, Clauset, Shalizi, and Newman (CSN) proposed another method based on the Kolmogorov-Smirnov (KS) statistics [18]. The estimated $x_{\min}$ is the value that minimizes the KS distance $D = \max_{x \geqslant x_{\min}} |F_n(x) - F(x)|$ between the empirical CDF and the CDF of the Pareto. Although CSN also show how to test the hypothesis that the data larger than $\hat{x}_{\min}$ are truly power-law distributed, their method provides only the best threshold, but does not tell whether, and how plausibly, different thresholds also determine a power-law tail.

Here we explore another approach, based on the ME distribution, which encompasses both the lognormal and the Pareto. The ME distribution is the result of the maximization of the Shannon's information entropy $W = \int -f(x) \log[f(x)]\, dx$ under constraints that impose the equality of the first k theoretical and empirical moments. The constraints are usually the arithmetic or geometric (characterizing) moments, respectively, given by $\int x^i f(x)\, dx$ and $\int \log(x)^i f(x)\, dx$, $i = 0, 1, \ldots, k$.

Let $\mu^i = \mathrm{E}[T(x)^i]$ and $\hat{\mu}^i = \frac{1}{n} \sum_j T(x_j)^i$ be, respectively, the ith theoretical and empirical characterizing moment. The ME approach entails maximizing W under the constraints $\mu^i = \hat{\mu}^i$ and can be solved introducing $k + 1$ Lagrange multipliers λ_i $(i = 0, \ldots, k)$, so that the solution (that is, the ME density) takes the form $f(x) = e^{-\sum_{i=0}^{k} \lambda_i T(x)^i}$. The Pareto distribution, $\mathrm{Par}(c,\alpha)$, is an ME density with $k = 1$, whereas the lognormal is ME with $k = 2$. For both distributions, the characterizing moments are the logarithmic ones. On the other hand, the exponential and the normal distributions are ME with $k = 1$ and 2, respectively, and arithmetic characterizing moments. The functions relating the parameters of the original and the ME distributions are detailed in Table I.[2]

[2]The functional forms of the relations among λ_0, λ_1, and λ_2 and the truncated normal parameters can be found in Ref. [41].

TABLE I. Parameters of the ME density for some commonly used distributions with logarithmic (log) and arithmetic (arithm) characterizing moments.

Distribution	Moments	λ_0	λ_1	λ_2
Pareto	log	$-\log(\alpha c^{\alpha})$	$\alpha+1$	–
Lognormal	log	$2\frac{\mu^2}{\sigma^2}-\frac{1}{2}\log\left(\frac{1}{2}\sigma^2\right)+\frac{1}{2}\log(\pi)$	$1-\frac{2\mu}{\sigma^2}$	$\frac{1}{2\sigma^2}$
Exponential	arithm	$-\log(\alpha)$	α	–
Normal	arithm	$\log(\sqrt{2\pi}\sigma)$	0	$\frac{1}{2\sigma^2}$

The most important issue in ME estimation is the choice of k. A larger number of constraints results in a more precise approximation, but also in a model with more parameters. Thus, the advantage of a better fit must be balanced against the noise caused by the estimation of more parameters. Accordingly, there are at least two ways of making a decision concerning the optimal value of k (e.g., k^*).

Since the maximized log-likelihood is equal to $-N\sum_{i=0}^{k}\lambda_i\hat{\mu}^i$ (N being the number of observations), we can compute a log-likelihood ratio (llr) test of the null hypothesis $k=k^*$ against $k=k^*+1$ as

$$\text{llr}=-2N\left(\sum_{i=0}^{k^*+1}\hat{\lambda}_i\hat{\mu}^i-\sum_{i=0}^{k^*}\hat{\lambda}_i\hat{\mu}^i\right).$$

Standard limiting theory guarantees that, asymptotically, the llr follows a χ^2_1 distribution and is optimal [42,48]. In this context optimality means that an llr with given size is uniformly at least as powerful as any other test with the same size, provided the size goes to zero sufficiently fast [49]. Thus the procedure is based on the following steps: (1) estimate sequentially the ME density with $k=1,2,\ldots$; (2) perform the test for each value of k; and (3) stop at the first value of k (e.g., k_0) such that the hypothesis $k=k_0$ cannot be rejected and conclude that $k^*=k_0$.

However, this method does not fully account for the costs of estimating a model with a larger number of parameters: This may introduce some further noise without substantially increasing the likelihood, and therefore the explanatory power, of the model. A common strategy to solve this problem [50,51] consists in computing an information criterion, such as the Akaike (AIC) or Bayesian (BIC) information criterion, which are still based on the maximized likelihood but introduce a penalization depending on the number of parameters. To avoid overfitting, one can then stop at the value k^* such that at least one of the following two conditions holds: (1) the llr test cannot reject the hypothesis $k=k^*$ or (2) the numerical value of AIC(k^*+1) [or BIC(k^*+1)] is larger than the numerical value of AIC(k^*) [or BIC(k^*)]. In the empirical analysis that follows we determine k^* by means of the combined used of the llr and the AIC when we apply ME estimation to the entire distribution of the data, whereas we use only the llr when focusing on the upper-tail behavior.

III. EMPIRICAL ANALYSIS

In recent years remarkable effort has been devoted to study the shape of the size distribution of cities [19,32–34]. The debate rests partly on the difficulty of properly defining what a city is and, empirically, what is the correct measure to employ [52]. By using data for all the populated places provided by the US Census in year 2000, it has been argued that the size distribution of cities is lognormal [33], not power-law as previously thought based on the largest metropolitan areas [32,53]. Yet, although the body of the city size distribution is well approximated by a lognormal, disagreement persists on whether there are significant departures in the upper tail [34,35]. The presence of a significant power-law tail has been recently confirmed—and the debate apparently closed—by means of the UMPU test [19]. We start our empirical analysis with an application of the UMPU, CSN, and ME tests to the same data on city size used in previous studies [19,33,34], so as to have a meaningful comparison of their relative performance.

Results are reported in Fig. 1. Depending on the test and the chosen significance level, we observe a power-law tail whose length ranges between top 536 and 1515 cities out of 25 359 populated places. In particular, according to the UMPU test the power law ranges between 1045 (10% level) and 1450 cities (1% level), the CSN test finds that the power-law tail is confined to the largest 536 cities, whereas according to the ME test, the power law spans 1205 (10%) to 1515 (1%) observations. For what concerns the shape parameter α, the classic Hill estimator is very sensitive to the choice of the threshold that marks the beginning of the power-law tail: Using the 10% significance level for the UMPU test ($x_{\min}=1045$), we obtain $\alpha=1.34$, which is not consistent with Zipf's law. The estimate of the shape parameter obtained using the $x_{\min}$ identified by CSN is 1.39, while the ME method yields a slightly smaller value (1.28).

However, before jumping to the conclusion that one theory or the other is supported by the data, one should take into account at least three related issues: (1) discriminating among the lognormal and the power-law tail behavior is difficult and the existing tests provide different results, (2) sample size matters, as well as truncations and other empirical phenomena that influence the estimation, and (3) the level of aggregation at which data are collected is not neutral to the detection of a power-law behavior. Thus we turn now to other real-world distributions at different levels of aggregation in order to better assess the influence of these various elements on the debate about the tail behavior in empirical data.

We analyze two widely investigated economic distributions: international trade flows [10,54–56] and business firm sizes [13,29–31]. First, we estimate the maximum entropy distribution against the (truncated) lognormal. Second, we analyze the behavior of the upper tail of the size distributions by means of the UMPU, CSN, and ME tests. Last, we discuss a theoretical model that properly account for the

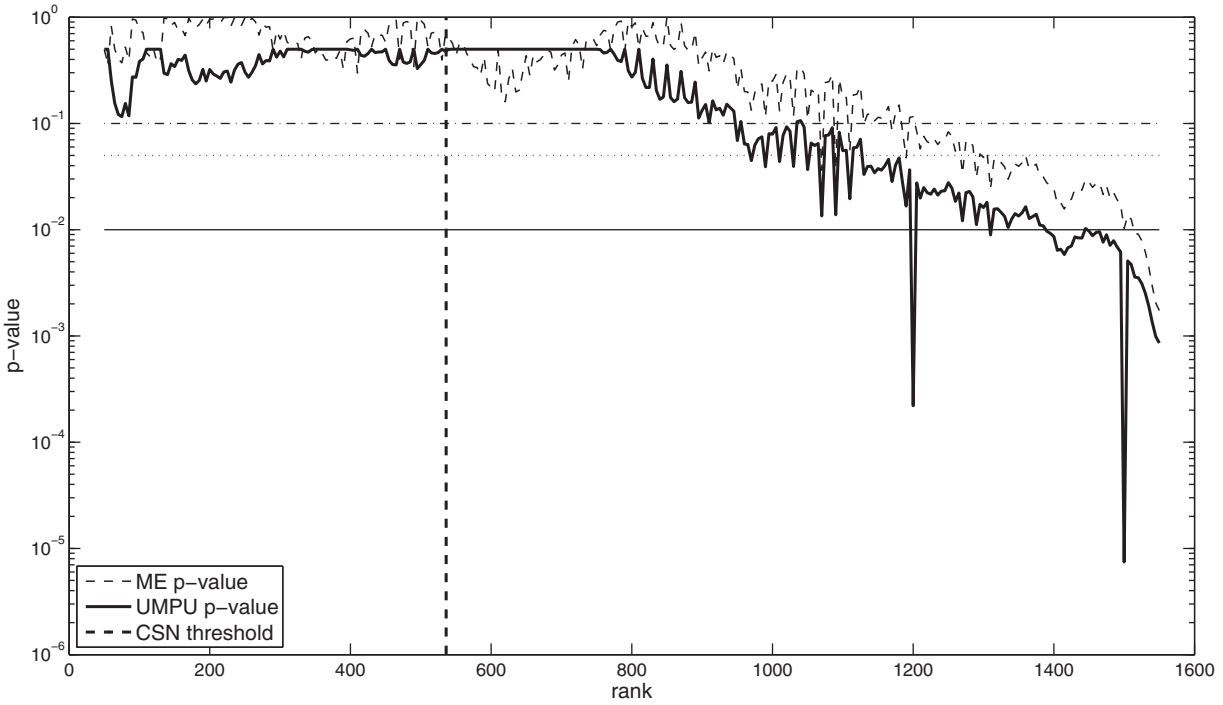

FIG. 1. (Natural log of) p-value of the UMPU and ME tests for exponential vs truncated normal distribution of the logarithm of US city sizes. The estimate of $x_{\min}$ obtained by CSN is also reported (vertical line).

emergence of a power-law tail in the firm size distribution [13,43].

It has often been noted that the Pareto tail of a system seems longer at a higher level of aggregation. For instance, simulating N observations from the same lognormal distribution, the top 10^2 observations look definitely Pareto when $N = 10^5$, and not Pareto when $N = 10^3$ [39]. Thus we analyze both trade and firm size at two levels of aggregation.

Trade data are taken from the COMTRADE database maintained by the United Nations. This collects data on 6 002 617 bilateral trade flows among 157 reporting countries (sources) and 230 destinations, at the six-digit level of the Harmonized System classification, which consists of roughly 5000 products. In the analysis we focus on the latest available year (2007) and analyze both disaggregate data at the level of single-product category and total trade obtained by summing up all trade flows for each of 20 767 nonnull country pairs.

To analyze the distribution of firm size we exploit a unique dataset on yearly sales of 916 036 pharmaceutical products by 5721 firms in 21 countries in 2004 [43,57]. Information is both available at the disaggregate level of product sales as well as reaggregated by assigning each product to the firm that sells it.

All data are expressed in thousands of US dollars. For notational convenience, in the following the original data and their logarithms will be called "levels" and "logarithms," respectively.

The distributions of both aggregate and disaggregate trade logarithms are not normal, but rather truncated normal, because of many small observations. Moreover, the observations smaller than zero (in logarithms) seem to be little informative, as there are clusters and peaks. Since (1) the distribution is truncated anyway, (2) the smallest observations do not appear very reliable, (3) there is a switch in the distribution approximately at $x_t = 1$ and (4) we are not particularly interested in the left tail of the distribution, we decided to discard the observations smaller than $x_t = 1$ and estimate just the left-truncated lognormal defined on $(x_t, +\infty)$. Notice that the distribution of all the observations available is likely to be a mixture of some distribution on $(0, x_t)$ and a left-truncated lognormal on $(x_t, +\infty)$.

As for the pharmaceutical data, all observations are larger than \$1000. However, it is clear that the distribution is truncated, in particular at the disaggregate level, so that we fit a truncated normal in this case as well, setting the truncation threshold equal to \$1000.

Figure 2(a) shows the distribution of the logarithms of the aggregate trade data with superimposed the optimal ME density with $k^* = 5$. For comparison purposes, the truncated normal density with parameters estimated from the data is shown as well. The ME density fits the data much better than the truncated normal. The fact that $k^* = 5$ also implies that the lognormal hypothesis for the levels should be rejected. For disaggregate trade data [Fig. 2(b)], the lognormal hypothesis is again rejected ($k^* = 6$), but the distance between the optimal ME and the normal seems smaller.

Turning now to the pharmaceutical data [Figs. 2(c) and 2(d)], the normal distribution is clearly not appropriate for the whole distribution at both aggregation levels. The distributions are more skewed than in the trade case, and the distance between the normal and the optimal ME looks large at both levels.

To compare more precisely the discrepancy between the density of a certain theoretical distribution and the

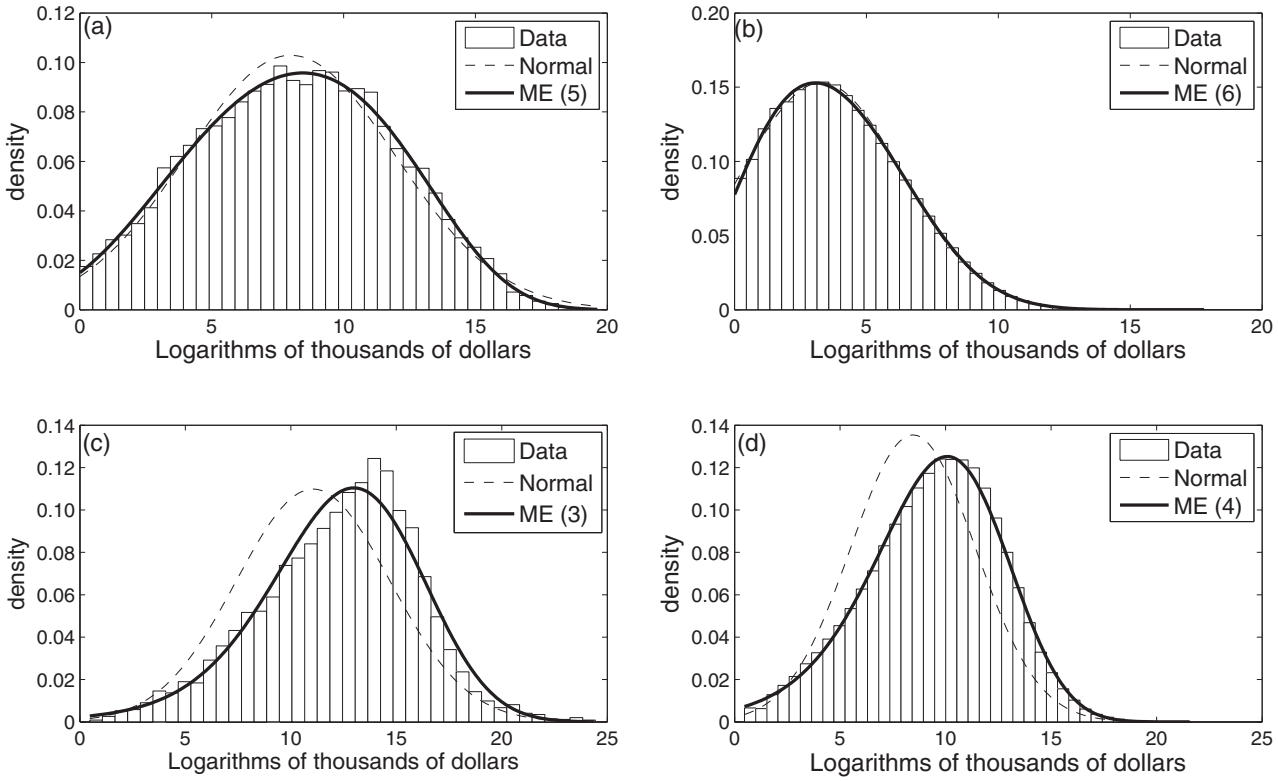

FIG. 2. Distributions of the logarithms of the data with superimposed the optimal ME density and the truncated normal density with parameters estimated from the data. From top to bottom and from left to right, the panels refer to aggregate trade data (a), disaggregate trade data (b), aggregate pharmaceutical data (c), and disaggregate pharmaceutical data (d).

ME, one can use the Kullback-Leibler distance $K(f||g) = \int f(x)\log(f(x)/g(x))dx$ or, more conveniently, the Information Distinguishability index $I_D(f||g) = 1 - e^{-K(f||g)}$, which is normalized to be included in the unit interval [58]. Table II reports the values of the two measures, where the MLEs of the truncated normal are obtained by means of the EM algorithm [59]. We find that the data offer greater support to the lognormal hypothesis for lower levels of aggregation, and this is particularly true in the case of trade.

We now focus on the tail behavior of the distributions. In order to estimate the threshold $x_{\min}$ (the Pareto scale parameter) we estimate the tail distribution for data $\{x : x > x_m\}$ for various x_m. For aggregate trade data, the threshold sequence goes from rank 50 to rank 800. We do not show results for ranks smaller than 50 (because estimates obtained with less than 50 observations are likely to be too unstable) and larger than 800 (because the Pareto hypothesis is definitely rejected by all tests for ranks larger than 800). According to the details in Sec. II, if the null hypothesis $k = 1$ cannot be rejected, the true distribution is Pareto.

TABLE II. Information Distinguishability index and Kullback-Leibler distance of the ME distribution from lognormal, for different levels of aggregation.

	Trade data		Pharmaceutical data	
	Aggregate	Disaggregate	Aggregate	Disaggregate
I_D index	0.0073	4.9702×10^{-4}	0.0711	0.0688
K distance	0.0073	4.9714×10^{-4}	0.0737	0.0713

Figure 3 shows the p-value of the llr test for $H_0 : k = 1$ against $H_1 : k = 2$ in the ME setup, the estimate of $x_{\min}$ obtained by CSN and the p-value of the UMPU test for exponential versus truncated normal [47]. The p-value of the power-law distribution found by CSN is reported in the caption.

In the case of aggregate trade [Fig. 3(a)] the evidence is mixed. At the 5% level, the Pareto hypothesis is valid approximately for ranks smaller than 650 (quantile 96.87%) according to the ME test, and for ranks smaller than 150 (quantile 99.28%) for the UMPU test. However, the results are far from clear-cut, as the p-value of the UMPU test is sometimes near 0.05 for ranks larger than 150. The CSN approach yields a rank equal to 408 (quantile 98.04%), but the p-value is equal to 0.024, so that the presence of a power-law tail seems to be questionable. The only conclusion that can be drawn with reasonable certainty is that the distribution is Pareto for ranks smaller than 150 and is not Pareto for ranks larger than 700 (quantile 96.63%). Note that, when we focus on the tail behavior, for ranks larger than 700 the ME procedure typically finds $k^* = 2$ (while rejecting $k = 1$), so that the distribution is a left-truncated lognormal in the upper tail.

Turning now to disaggregate trade [Fig. 3(b)] the Pareto hypothesis is valid approximately for ranks smaller than 1600 for ME (quantile 99.97%), 1480 for CSN (quantile 99.98%), and 300 for UMPU (quantile >99.99%). However, similarly to the case of aggregate data, for ranks between 500 (>99.99%)

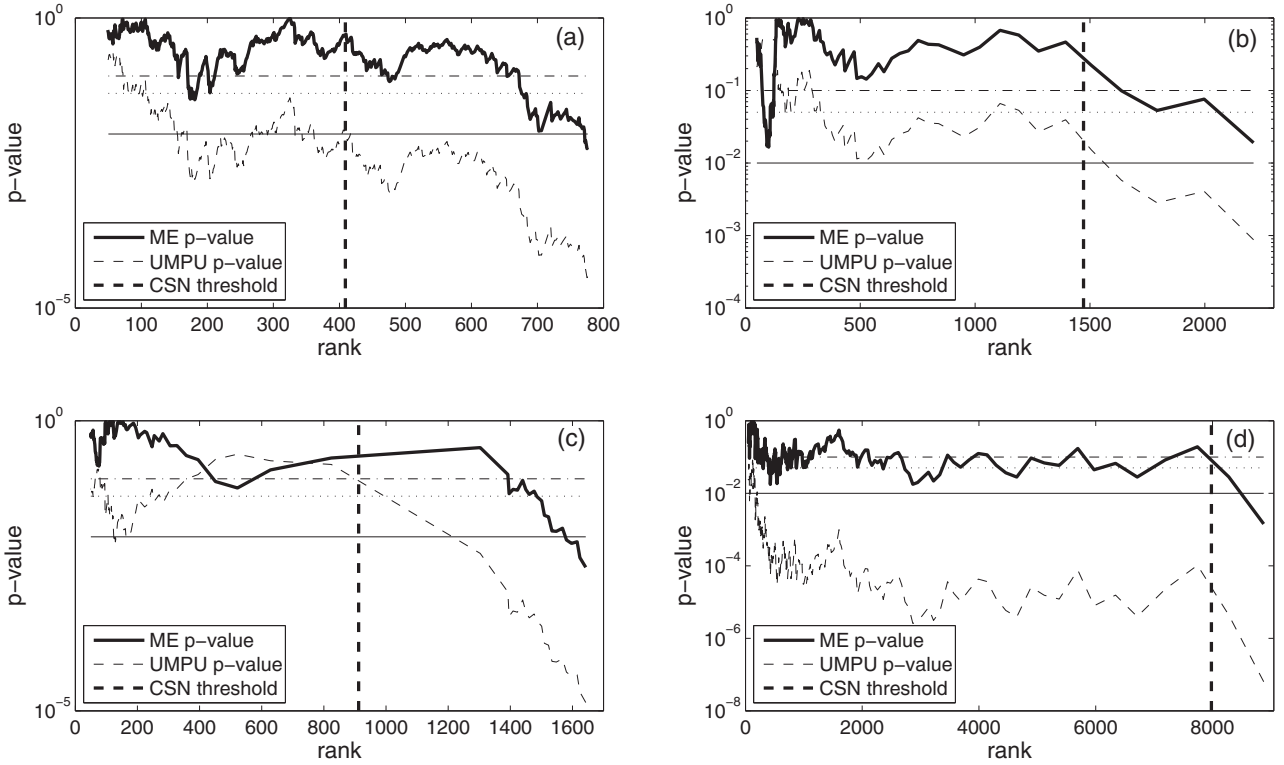

FIG. 3. The graphs show the (natural log of) p-value of the llr test for $H_0 : k = 1$ against $H_1 : k = 2$ in the ME setup, the estimate of $x_{\min}$ obtained by CSN and the p-value of the UMPU test for exponential versus truncated normal. From top to bottom and from left to right, the panels refer to aggregate trade data (a), disaggregate trade data (b), aggregate pharmaceutical data (c), and disaggregate pharmaceutical data (d). The p-value of the CSN threshold is equal to 0.024 for aggregate trade, 0.274 for disaggregate trade, 0.026 for aggregate pharmaceutical, and 0.868 for disaggregate pharmaceutical data.

and 1500 (quantile 99.98%), the p-value of the UMPU test is sometimes near 0.05. The p-value of the CSN approach (0.274) seems to confirm that the distribution is power-law, and all tests suggest that the distribution is not Pareto for ranks larger than 1600. Although the ranks such that the power-law hypothesis is accepted are larger for disaggregate data, the population size is much larger, so that only a very small fraction of the observations is generated by a Pareto tail.

As for pharmaceutical data, Fig. 3(c) shows that at the 5% level the distribution is Pareto for ranks approximately smaller than 1500 for ME (quantile 73.78%). The length of the Pareto tail found by UMPU is approximately 1000 (quantile 82.52%), while CSN stops at rank 900 (quantile 84.27%). The p-value of CSN is small (0.026), so that the tail may actually not be Pareto. Notice that for aggregate data the p-value of the UMPU test is not always below the one corresponding to ME, and the latter first goes under the 5% level at rank 400. Finally, for what concerns disaggregated figures [Fig. 3(d)], the UMPU test starts rejecting the null hypothesis of Pareto very early, for ranks approximately equal to 300 (quantile 99.97%). On the contrary, the CSN approach identifies a much longer power-law tail, roughly corresponding to the largest 8000 observations (quantile 99.13%). The corresponding p-value is large (0.868), suggesting that the upper tail of the distribution is likely to actually be Pareto. The ME test gives a similar picture, as it starts staying definitively below the 5% level at rank near 8000. However, starting at rank 2500 the p-value fluctuates between 0.10 and 0.01, making it difficult to draw a clear-cut conclusion. Disaggregate pharmaceutical data thus represent a clear example that the three procedures may yield different results. Moreover, in this case it is not clear whether the early rejection of the Pareto hypothesis by the UMPU test implies a good performance or not.

One of the benefits of the ME approach is that it delivers the estimated parameters of the distribution. Table III shows the estimates of the Pareto shape parameter α by means of the different approaches. UMPU and CSN both rely on the Hill estimator: The difference in the estimated coefficient stems from the fact that they identify two different thresholds for the beginning of the Pareto tail, and the Hill estimator is quite sensitive to this. We always have $\hat{\alpha}^{(\mathrm{UMPU})} > \hat{\alpha}^{(\mathrm{CSN})} > \hat{\alpha}^{(\mathrm{ME})}$, and the values estimated by the three methodologies are not always very close to each other. This is especially true for $\hat{\alpha}^{(\mathrm{UMPU})}$ and $\hat{\alpha}^{(\mathrm{ME})}$: The difference is particularly large in the case of disaggregate pharmaceutical data.

TABLE III. Estimates of the Pareto shape parameter. $\hat{\alpha}^{(\mathrm{CSN})}$ is the estimated parameter generated by the CSN method minus 1.

	Trade data		Pharmaceutical data	
	Aggregate	Disaggregate	Aggregate	Disaggregate
$\hat{\alpha}^{(\mathrm{ME})}$	0.948	1.380	0.532	1.021
$\hat{\alpha}^{(\mathrm{CSN})}$	1.080	1.402	0.601	1.038
$\hat{\alpha}^{(\mathrm{UMPU})}$	1.190	1.551	0.623	1.513

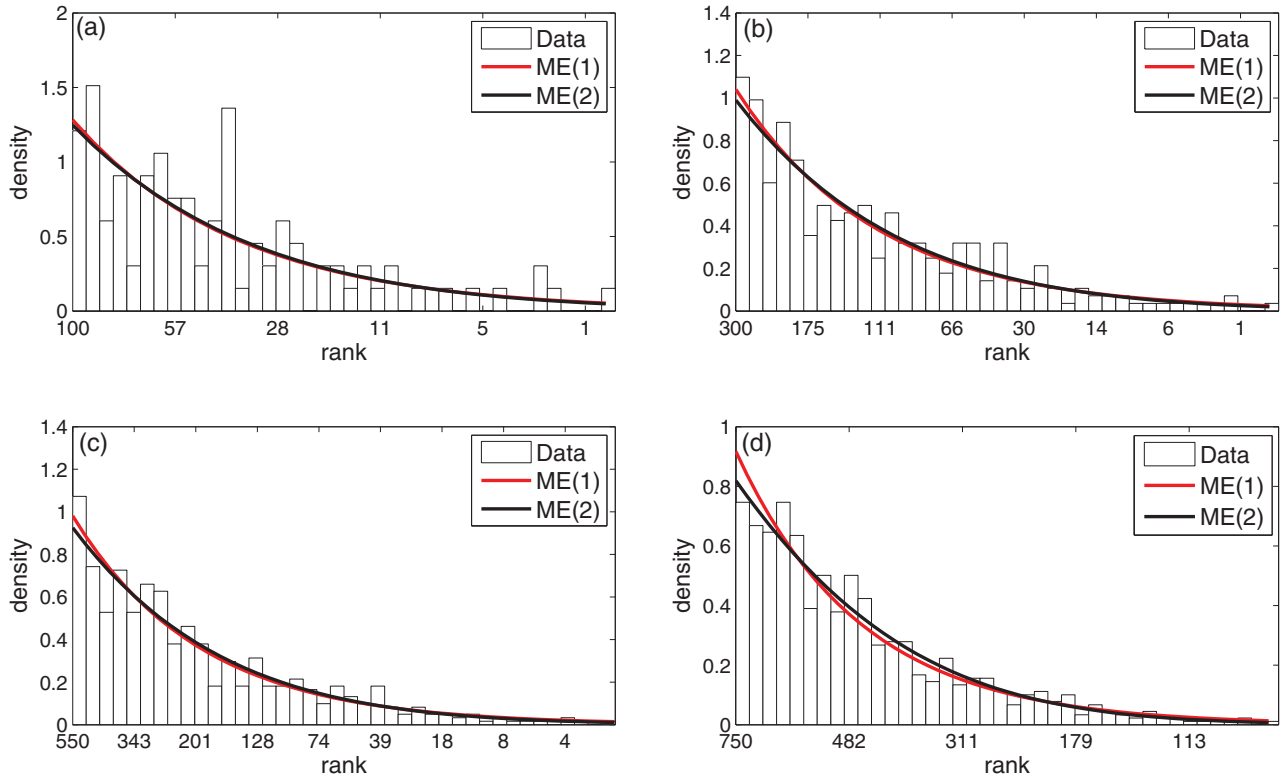

FIG. 4. (Color online) Results for the sampled aggregate trade data. ME(1) and ME(2) are the exponential and the truncated normal distribution, respectively.

In some cases the evidence from the three tests is not the same. It may therefore be of interest to check how different the ME densities are when $k = 1$ (Pareto) and $k = 2$ (truncated lognormal). To this aim, Fig. 4 gives some insights for the aggregate trade data. The graph displays the histogram of the logarithms above four different thresholds and the estimated ME(1) and ME(2) densities (respectively exponential and truncated normal when using the logarithms). The four thresholds correspond to ranks in different positions of the tails: in particular, panel (a) uses rank 100, such that all tests accept the Pareto hypothesis, panels (b) and (c) use rank 300 and 550, such that the UMPU test rejects but the ME test accepts the Pareto hypothesis, and finally panel (d) uses rank 750, for which all tests reject the Pareto hypothesis.

It can be seen that the two densities are almost indistinguishable for the three smallest ranks, for which the tests give somewhat different results. On the other hand, when the rank is equal to 750 and all the tests suggest rejection of the Pareto hypothesis, the difference is more evident. These results are quite reassuring, as they show that, when the outcomes of the tests are not unambiguous, the possible data-generating processes are almost identical.

An issue that requires further investigation is the following. As pointed out in Sec. II, under the lognormal hypothesis, when the threshold is large, so that the number of observations is small, it is often observed that the tail seems to follow a Pareto distribution. In order to quantify how the sample size influences the statistical features of the tail (and, in particular, the estimated $x_{\min}$), we apply again the tests to a sample of the disaggregate data of the same size as the aggregate populations ($n = 20\,767$ for trade and $n = 5721$ for pharmaceutical data). Thus, the size of the two datasets to be compared is now the same. The results are reported in Fig. 5.

In qualitative terms, for the trade data, the outcome is similar to the one shown in Fig. 3(d); the thresholds obtained with the three tests are also in good agreement with each other. The length of the Pareto tail in the sampled data ranges between quantiles 98.56% (rank 300, CSN) and 97.11% (rank 600, ME), compared with quantile 99.98% in the original data. For the pharmaceutical data, the Pareto tail reaches approximately quantiles 95.63% (rank 250, CSN) or 94.06% (rank 340 ME) in the sampled data, whereas it is confined above quantile 99% in the original data.

Consistent with the observation that the sample size affects the length of the Pareto tail observed in the data [39], the Pareto tail is more pronounced in the samples than in the original population. In particular, the difference between the length of the Pareto tail in the population and in the sample is larger for the trade than for the pharmaceutical sales data, as was to be expected in view of the larger difference between the population and the sample size in the case of trade data.

Disaggregate data show, in the best case, a power-law tail confined to the last percentile of the distributions. However, trade and pharmaceutical data show a different behavior upon aggregation. Since the aggregate trade distribution is certainly not Pareto below the 96.63% percentile and a sample of the same size of the disaggregate trade distribution has a Pareto tail starting at the quantile 97.5%, we can conclude that the trade distribution has a very short Pareto tail (if any). Conversely, the business firm distribution has a power-law tail from the 75.53%

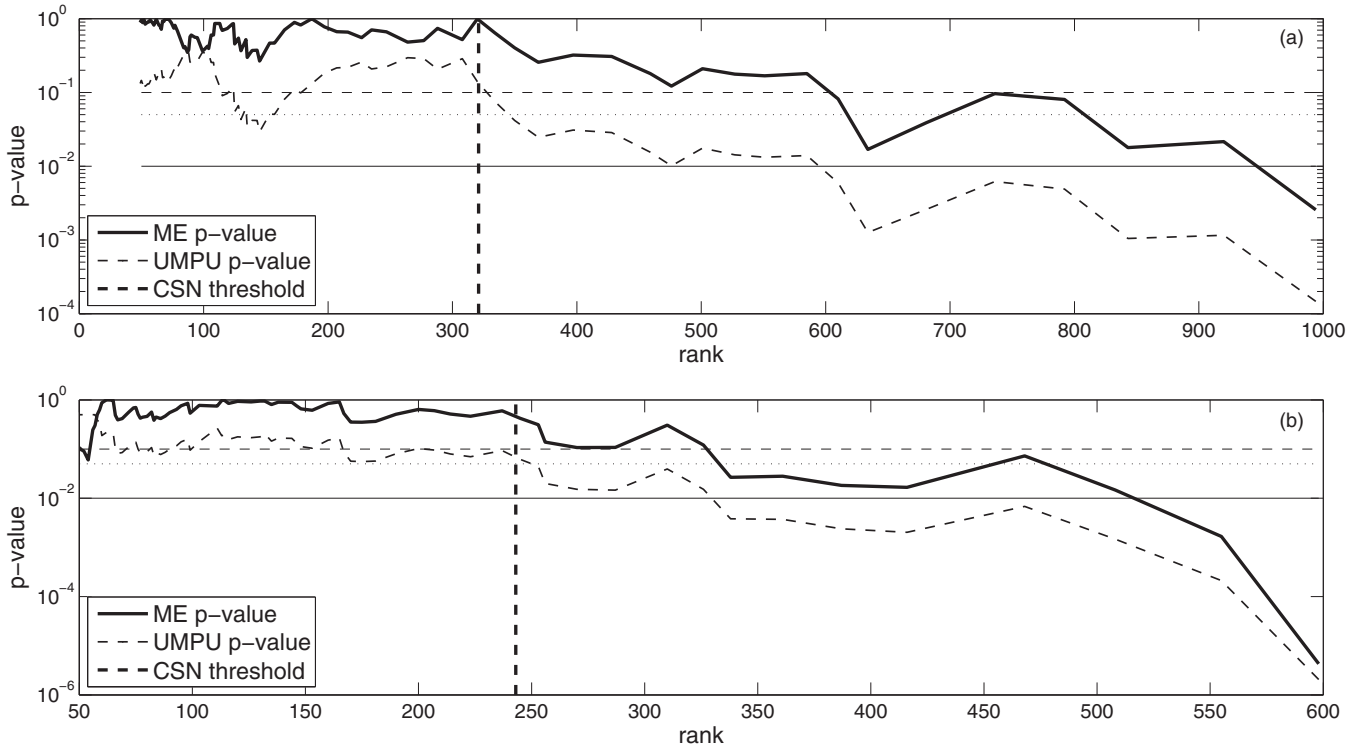

FIG. 5. Results for the length of the Pareto tail for the sampled trade (a) and pharmaceutical (b) data. The p-value of the Pareto tail found by means of the CSN approach is 0.096 for trade and 0.084 for pharmaceutical data.

quantile onward, while the Pareto tail of a sample of the same size from the disaggregate (pharmaceutical) data is limited to the quantile 94.1% (at most). Therefore the Pareto tail of business firm size distribution emerges upon aggregation, not just as a matter of sample size.

To make sense of this, it has been argued elsewhere [13] that the power-law tail of the firm size distribution can be generated as a sum of lognormals. In fact, the aggregate size of each firm (S) is given by the sum of the size of products (s) over the total number of products sold (K). In this context the firm size distribution can be usefully approximated by a lognormal distribution multiplied by a stretching factor which increases with S.

Figure 6 shows that the complementary CDF of the number of products sold by each pharmaceutical firms $P(K_f)$ is approximately Pareto, whereas the same distribution for international trade $P(K_c)$ (number of products traded by each country pair) is far less skewed. Thus the emergence of a power-law tail in the pharmaceutical data can be explained by the presence of a Pareto component in the stretching factor [13].

To substantiate the claim that the Pareto tail in pharmaceutical data can be generated by the aggregation of products into firms according to a very skewed distribution $P(K_f)$, we run the tests on a synthetic dataset obtained by aggregating product-level data according to $P(K_c)$ instead of $P(K_f)$. We find that the Pareto tail is limited to ranks ranging from 136 (CSN) to 162 (ME test), which correspond to quantiles between 97.55% and 97.17%. Hence, the power-law tail is much smaller than in the true aggregate dataset, a result in line with the conjecture that the skewness of the aggregation rule $P(K_f)$ contributes to the emergence of a Pareto tail in the data.

IV. COMPARING DIFFERENT TESTS VIA SIMULATIONS

The empirical analysis conducted in Sec. III shows that the results of the UMPU, ME, and CSN procedures are different. Yet without knowing *ex ante* where the true threshold $x_{\min}$ lies, we cannot assess whether an earlier rejection of the null hypothesis of a Pareto tail actually represents a desirable feature of the test or, on the other hand, a more powerful test should identify a longer power-law tail as claimed in the case of city size [19]. To better assess the relative merits of the three test employed in the paper we turn now to simulation analysis.

In particular, we simulate 25 000 observations from a mixture of a right-truncated lognormal and a Pareto distribution, with density given by

$$f(x) = \begin{cases} r\dfrac{1}{\Phi\left(\frac{\log(\mu)-x_{\min}}{\sigma}\right)} f_1(x), & x \leqslant x_{\min}, \\ (1-r) f_2(x), & x > x_{\min}, \end{cases} \tag{2}$$

where Φ is the CDF of the normal distribution, f_1 and f_2 are the Logn(μ,σ^2) and the Par$(x_{\min},\alpha)$ densities. For the density to be continuous and differentiable at $x_{\min}$ we need to impose some restrictions on the lognormal expected value μ and the mixing weight r, which are not free parameters, and are given by [60]

$$\mu = \log(x_{\min}) - \alpha\sigma^2; \quad r = \frac{\sqrt{2\pi}\alpha\sigma\,\Phi(\alpha\sigma)e^{\frac{1}{2}(\alpha\sigma)^2}}{\sqrt{2\pi}\alpha\sigma\,\Phi(\alpha\sigma)e^{\frac{1}{2}(\alpha\sigma)^2}+1}.$$

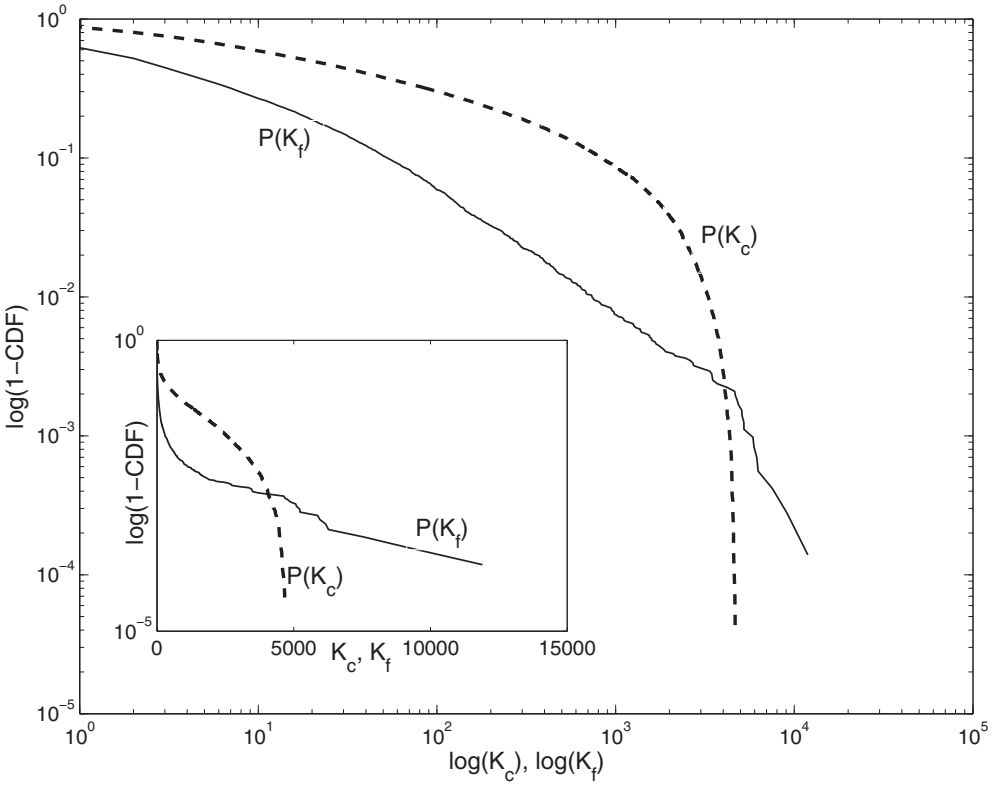

FIG. 6. The complementary cumulative distribution of the number of commodities traded by country pairs $P(K_c)$(trade data, broken line) and products by firm $P(K_f)$ (pharmaceutical data, solid line). Double logarithmic scale (main figure) and semilogarithimic scale (inset).

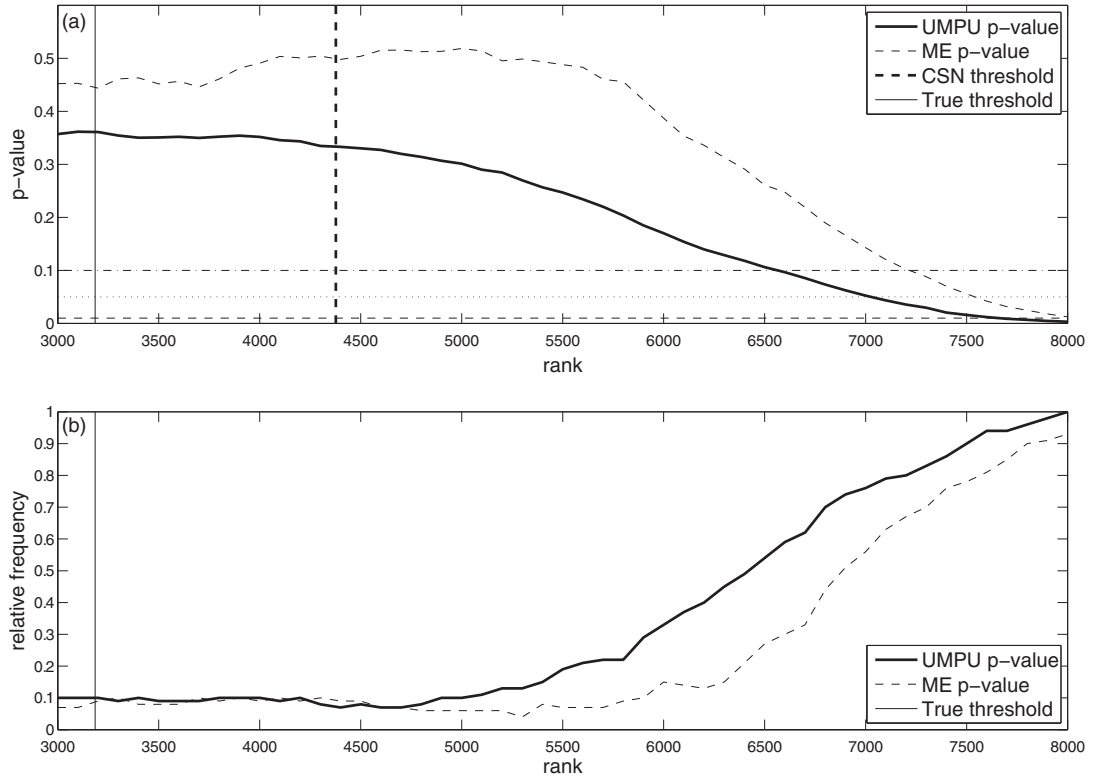

FIG. 7. (a) Average p-value (across 100 replications) for the ME and UMPU tests applied to the lognormal-Pareto mixture defined in (2). The dashed line represents the average $x_{\min}$ identified by the CSN approach. (b) Relative frequency of replications for which the ME and UMPU tests reject the null hypothesis of Pareto for different thresholds.

In the simulation we set $\sigma = 1$, $x_{\min} = 20$, $\alpha = 1.3$, giving $\mu = 1.696$ and $r = 0.8726$. With $N = 25\,000$, this implies that there are 3185 observations larger than $x_{\min}$ (i.e., simulated from the Pareto distribution). Note also that the probability that a Logn(1.696, 1) distribution is smaller than 20 is equal to 0.903: thus, in the mixture, the smallest 90.3% observations come from the lognormal, and therefore are certainly not yet Gumbel, in the sense that they are not yet far enough in the tail for the BHP theorem to work.

We implement the UMPU, CSN, and ME tests on the simulated sample: Fig. 7(a) shows the average p-value across 100 replications (UMPU and ME) together with the average $x_{\min}$ delivered by CSN. The graph shows that in this case CSN performs better than both UMPU and ME, in that it identifies (on average) a shorter power-law tail in the data, closer to the true threshold. Panel (b) of Fig. 7 displays the empirical power function of the simulated data for UMPU and ME, i.e., the number of times each test rejects the null hypothesis at the 5% level divided by the number of replications. The graph confirms that UMPU is more powerful than ME, although both tests find a much longer Pareto tail than the true one and should therefore be used with this caveat in mind.

Our simulations show that the three tests find power-law tails whose lengths rank as in the empirical data on city size, i.e., CSN, UMPU, and ME. Furthermore all tests find Pareto tails that are longer that the true value. A possible interpretation of our results is that in the case of the city size distribution, the length of the power-law tail could actually be shorter than what is currently thought, with the value found by CSN acting as an upper bound, and a shape parameter α even further away from what Zipf's law would imply. In any case, for what concerns cities, the power-law tail is much shorter than that observed for pharmaceutical firms, though slightly longer than the one we find for trade data. Hence, if an amplification mechanisms is actually at play, it is limited to the largest cities.[3]

Although further analysis on the relative performance of the various tests is probably necessary, our simulation exercise questions the existence of a clear-cut ranking in the three tests analyzed in the paper and reinforces the impression that in applied research one should avoid relying on a single method to identify the existence and the length of a Pareto upper tail in the data.

ACKNOWLEDGMENTS

The authors would like to thank two anonymous referees for their useful insights. Furthermore, they received helpful comments from Jakub Growiec, Alex Petersen, Michael F. Shlesinger, and Emanuele Taufer. Responsibility for all remaining errors rests solely with the authors.

[3]We have no means to directly test the effect of aggregation on the distribution of city size. However, a recent work that uses a clustering algorithm to define cities finds that the 1947 clusters whose population is larger than 12 000 are well approximated by a power law [52]. By applying the same testing procedure to the data on populated places, we find a Pareto tail spanning just 17 cities. As long as clusters are combinations of populated places, this evidence is consistent with the idea that the level of aggregation at which data are analyzed matters.

[1] H. A. Simon, Biometrika **42**, 425 (1955).
[2] E. Montroll and M. Shlesinger, Proc. Natl. Acad. Sci. USA **79**, 3380 (1982).
[3] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[4] J. Brown, V. Gupta, B. Li, B. Milne, C. Restrepo, and G. West, Philos. Trans. R. Soc. London B **357**, 619 (2002).
[5] W. J. Reed and B. D. Hughes, Phys. Rev. E **66**, 067103 (2002).
[6] M. Mitzenmacher, Internet Math. **1**, 226 (2004).
[7] D. Sornette, *Critical Phenomena in Natural Sciences: Chaos, Fractals, Self-organization, and Disorder: Concepts and Tools* (Springer, Heidelberg, 2004).
[8] M. Newman, Contemp. Phys. **46**, 323 (2005).
[9] M. A. Reid and C. D. Wilson, Astrophys. J. **650**, 970 (2006).
[10] X. Gabaix, Annu. Rev. Econ. **1**, 255 (2009).
[11] Y. Y. Kagan, Tectonophysics **490**, 103 (2010).
[12] A. Allen, B. Li, and E. Charnov, Ecol. Lett. **4**, 1 (2001).
[13] J. Growiec, F. Pammolli, M. Riccaboni, and H. E. Stanley, Econ. Lett. **98**, 207 (2008).
[14] M. Goldstein, S. Morris, and G. Yen, Eur. Phys. J. B **41**, 255 (2004).
[15] X. Gabaix and Y. M. Ioannides, in *Handbook of Regional and Urban Economics*, 1st ed., edited by J. V. Henderson and J. F. Thisse, Vol. 4, Chap. 53 (Elsevier, Amsterdam, 2004), pp. 2341–2378.
[16] H. F. Coronel-Brizio and A. R. Hernández-Montoya, Physica A: Stat. Mech. Appl. **354**, 437 (2005).
[17] V. F. Pisarenko and D. Sornette, Physica A: Stat. Mech. Appl. **366**, 387 (2006).
[18] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Rev. **51**, 661 (2009).
[19] Y. Malevergne, V. Pisarenko, and D. Sornette, Phys. Rev. E **83**, 036111 (2011).
[20] R. Fisher, A. Corbet, and C. Williams, J. Anim. Ecol. **12**, 42 (1943).
[21] S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton University Press, Princeton, 2001).
[22] W. Reed, Econ. Lett. **74**, 15 (2001).
[23] Z. Cheng and S. Redner, Phys. Rev. Lett. **60**, 2450 (1988).
[24] B. McBreen, K. J. Hurley, R. Long, and L. Metcalfe, Mon. Not. R. Astron. Soc. **271**, 662 (1994).
[25] J. Söderlund, L. B. Kiss, G. A. Niklasson, and C. G. Granqvist, Phys. Rev. Lett. **80**, 2386 (1998).
[26] V. Abramenko and D. Longcope, Astrophys. J. **619**, 1160 (2005).
[27] V. Khvorostyanov and J. Curry, J. Geophys. Res. **111**, D12202 (2006).
[28] L. Benguigui and E. Blumenfeld-Lieberthal, Int. J. Mod. Phys. C **17**, 1429 (2006).

[29] Y. Ijiri and H. Simon, *Skew Distributions and the Sizes of Business Firms* (North-Holland, Amsterdam, 1977).
[30] J. Sutton, J. Econ. Litt. **35**, 40 (1997).
[31] R. L. Axtell, Science **293**, 1818 (2001).
[32] X. Gabaix, Q. J. Econ. **114**, 739 (1999).
[33] J. Eeckhout, Am. Econ. Rev. **94**, 1429 (2004).
[34] M. Levy, Am. Econ. Rev. **99**, 1672 (2009).
[35] J. Eeckhout, Am. Econ. Rev. **99**, 1676 (2009).
[36] M. Williamson and K. J. Gaston, J. Anim. Ecol. **74**, 409 (2005).
[37] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance* (Springer, Heidelberg, 1997).
[38] G. De Fabritiis, F. Pammolli, and M. Riccaboni, Physica A: Stat. Mech. Appl. **324**, 38 (2003).
[39] R. Perline, Stat. Sci. **20**, 68 (2005).
[40] R. Hisano and T. Mizuno, Physica A: Stat. Mech. Appl. (Amsterdam) **390**, 309 (2011).
[41] J. Kapur, *Maximum Entropy Models in Science and Engineering* (Wiley, New York, 1989).
[42] X. Wu, J. Econ. **115**, 347 (2003).
[43] D. Fu, F. Pammolli, S. Buldyrev, M. Riccaboni, K. Matia, K. Yamasaki, and H. Stanley, Proc. Nat. Acad. Sci. USA **102**, 18801 (2005).
[44] R. Rubinstein and D. Kroese, *The Cross-Entropy Method* (Springer, Heidelberg, 2004).
[45] C. Kleiber and S. Kotz, *Statistical Size Distributions in Economics and Actuarial Sciences* (Wiley, New York, 2003).
[46] B. M. Hill, Ann. Stati. **3**, 1163 (1975).
[47] J. del Castillo and P. Puig, J. Am. Stat. Assoc. **94**, 529 (1999).
[48] D. Cox and D. Hinkley, *Theoretical Statistics* (Chapman and Hall, New York, 1974).
[49] R. Serfling, *Approximation Theorems of Mathematical Statistics* (Wiley, New York, 1980).
[50] W. H. Greene, *Econometric Analysis* (Prentice Hall, Upper Saddle River, 2003).
[51] G. Koop, *Bayesian Econometrics* (Wiley, New York, 2003).
[52] H. D. Rozenfeld, D. Rybski, X. Gabaix, and H. A. Makse, Am. Econ. Rev. (to be published).
[53] G. K. Zipf, *Human Behavior and the Principle of Last Effort* (Addison-Wesley, Cambridge, MA, 1949).
[54] K. Bhattacharya, G. Mukherjee, J. Saramäki, K. Kaski, and S. S. Manna, J. Stat. Mech. (2008) P02002.
[55] W. Easterly, A. Reshef, and J. Schwenkenberg, *The Power of Exports*, Policy Research Working Paper Series 5081 (World Bank, Washington, DC, 2009).
[56] M. Riccaboni and S. Schiavo, New J. Phys. **12**, 023003 (2010).
[57] M. Riccaboni, F. Pammolli, S. V. Buldyrev, L. Ponta, and H. E. Stanley, Proc. Natl. Acad. Sci. USA **105**, 19595 (2008).
[58] E. Soofi, N. Ebrahimi, and M. Habibullah, J. Am. Stat. Assoc. **90**, 657 (1995).
[59] A. P. Dempster, N. M. Laird, and D. B. Rubin, J. R. Stat. Soc. Ser. B (Methodological) **39**, 1 (1977).
[60] D. P. M. Scollnik, Scand. Actuar. J. **1**, 20 (2007).