# Where Gibrat meets Zipf: Scale and Scope of French Firms

Marco Bee* Massimo Riccaboni† Stefano Schiavo‡

February 14, 2017

**Abstract**

The proper characterization of the size distribution and growth of firms represents an important issue in economics and business. We use the Maximum Entropy approach to assess the plausibility of the assumption that firm size follows Lognormal or Pareto distributions, which underlies most recent works on the subject. A comprehensive dataset covering the universe of French firms allows us to draw two major conclusions. First, the Pareto hypothesis for the whole distribution should be rejected. Second, by discriminating across firms based on the number of products sold and markets served, we find that, within the class of multi-product companies active in multiple markets, the distribution converges to a Zipf's law. Conversely, Lognormal distribution is a good benchmark for small single-product firms. The size distribution of firms largely depends on firms' diversification patterns.

**Keywords:** Firm size distribution; Firm diversification; Pareto distribution; Zipf's law; International trade

**JEL Codes:** C46, L11, L25

*Department of Economics and Management, University of Trento, Via Inama, 5 - 38122 Trento

†LIME, IMT School for Advanced Studies, Piazza S. Francesco, 19 - 55100 Lucca; Department of Managerial Economics, Strategy and Innovation (MSI), KU Leuven Naamsestraat 69 - 3000 Leuven

‡Department of Economics and Management, University of Trento, Via Inama, 5 - 38122 Trento

# 1 Introduction

At least since the work of Gibrat (1931), a great deal of attention has been devoted to the investigation of the size distribution of business firms and the mechanics through which it is determined (Luttmer, 2011). In his seminal work, Gibrat (1931) found that French *establishments* were lognormally distributed, and postulated a process of proportional growth capable of replicating that shape. A few decades later, Simon and Bonini (1958) showed that a Pareto distribution provides a better fit to the upper tail of the *largest firms*. Before jumping to the conclusion that Gibrat and Simon identified different benchmarks for the size of firms, it should be noticed that the former analyzed *all* establishments while the latter focused on the largest US companies. Since then, the lognormal and the Pareto distributions have been considered to be the two foremost candidate distributions for firm size (see Sutton, 1997; Mitzenmacher, 2004; de Wit, 2005, for broad reviews of the topic). More recently, a special case of the Pareto distribution, namely Zipf's law, has come to be considered the best first-cut benchmark for firm size and many other empirical phenomena (Simon, 1955; Axtell, 2001; Luttmer, 2007; Gabaix, 2009) and it is increasingly used also in economic theory to model, for instance, firm heterogeneity (Melitz, 2003). Even though the Zipf's law is becoming more and more popular as an analytically convenient assumption in economic modeling, on the empirical ground there is no conclusive evidence in favor of a power-law shape of the distribution of firm sizes.

Already in the Seventies, Ijiri and Simon (1974) noticed that the data feature systematic departures from power-law behavior, as most firm size distributions depict a concave shape in a double logarithmic scale graph. Due to the lack of data on the whole distribution of firms, early works analyzed only the largest (listed) companies (such as Fortune 500 firms, see for instance Hart and Prais, 1956; Simon and Bonini, 1958). Axtell (2001), who represents the main reference of the majority of recent papers, manages to overcome the limitation in the representativeness of the data by using information taken from the US Census, thus looking

at the entire distribution of firms. He concludes that Zipf's law provides the best description of the data "all the way down to the smallest sizes" (Axtell, 2001, p. 1819), even though (as noted by Axtell himself) a closer inspection reveals the presence of departures similar to those highlighted by Ijiri and Simon (1974), compatible with a lognormal distribution. The concave shape of the size distribution has been further confirmed by Rossi-Hansberg and Wright (2007), who use the same data analyzed by Axtell (2001). This is part of a stream of literature that claims there is no universal functional form for the size distribution of firms, as it depends on specific industry characteristics (Kumar et al, 1999; Halvarsson, 2013) such as innovation (Klette and Kortum, 2004), financial constraints (Cabral and Mata, 2003), learning and firm selection (Jovanovic, 1982; Luttmer, 2007), international trade (Di Giovanni et al, 2011), institutional factors (Garicano et al, 2013) and the level of aggregation (Tang, 2015).

The typical strategy adopted in this literature considers one candidate distribution (Zipf, Pareto or, less frequently, lognormal) and performs goodness of fit analyses. Replicating existing results is often difficult because of the limitations in accessing official data, and only a few studies investigate the universe of firms in the US (Axtell, 2001; Rossi-Hansberg and Wright, 2007; Luttmer, 2011) or other countries (see Cabral and Mata, 2003 for Portugal, Eaton et al, 2011 or Garicano et al, 2013 for France and Huber and Pfaffermayr, 2010 for Austria).

The contribution of this paper is twofold. First, we take a closer look at the Pareto hypothesis, with a degree of methodological accuracy that goes beyond the existing literature. In this enterprise, we exploit the massive information content of a database covering virtually all French firms: this is particularly important as discriminating between a lognormal and a Pareto upper tail is extremely difficult from a statistical point of view (Malevergne et al, 2009; Bee et al, 2011). Moreover, it is crucial to cover all firms down to the smallest sizes. The empirical study of the characteristics of the (tail of the) distribution is performed by means of different methods, in order to be able to identify possible pitfalls of single approaches

related to specific features of the data under analysis. In addition to the tests, we employ the Maximum Entropy (ME) method to find the best approximating density in a non-parametric setup. Furthermore, consistently with the analysis by Virkar and Clauset (2014), our results confirm that the practice of binning the data, quite common in the literature on the distribution of firm size, significantly affects the results and should be avoided. We find that the whole data distribution is neither lognormal nor Pareto, thus suggesting that the debate in the literature about the shape of the size distribution is partially misleading.

Second, we argue that it is far more interesting to decompose the population of firms into groups with different characteristics to see if some of them exhibit a more pronounced Pareto or lognormal behavior. In this paper we discriminate across firms based on their scope, e.g. number of products and markets they serve. The chance to operate such a classification represents a further advantage offered by the data we use. We find that the size distribution of firms largely depends on firms' diversification patterns. In particular, our analysis shows that the size distribution changes according to the features of this classification and its upper tail converges to a Pareto distribution with a shape parameter approaching 1 from above (Zipf's law) for highly diversified and multinational companies whereas the lognormal distribution is a good benchmark for small and mid-sized enterprises. Our work contributes to recent attempts to reinvigorate research on firms' diversification as a crucial aspect of the growth-of-firms (Coad, 2009).

The rest of the paper is organized as follows: the next section addresses the most important methodological issues relative to the identification of a power-law behavior and describes the data. Section 3 presents the main empirical outcomes and investigates the size distribution of different groups of firms. Finally, Section 4 discusses the implications of our findings for economic theory and offers some concluding remarks.

# 2 Methodology Background and Data

The empirical literature dealing with firm size distribution is characterized by a number of often overlooked methodological issues that may have a significant impact on the results. The most important ones concern the power of the tests used to distinguish among different candidate distributions, especially when the sample size is small: the widespread practice of binning the data and whether the power-law behavior refers to the upper tail or to the whole distribution.

A substantial body of literature addresses the problem of discriminating between power-law (Pareto) and lognormal tail behavior. The two distributions are mathematically different, but only in the limit (Perline, 2005), so that for finite sample size the tests often have low power. Given these premises, several tests have been developed. Here we follow Bee et al (2013) and show the results obtained with the Uniformly Most Powerful Unbiased (UMPU) test developed by Del Castillo and Puig (1999) and used by Malevergne et al (2011); the Maximum Entropy (ME) test by Bee et al (2011); the test proposed by Gabaix and Ibragimov (2011, GI henceforth).

The size distribution of firms has often been tested using binned data (see for instance Axtell, 2001; Di Giovanni et al, 2011). Axtell (2001) uses US firm sizes measured by receipts in dollars (US Census Bureau data for 1997, consisting in 5 541 918 observations). Data are tabulated in successive bins of increasing size in powers of three, so that bins are equally spaced in logarithmic scale. Using the 7 bins obtained in this way, an OLS regression in doubly logarithmic scale suggests an approximate Zipf distribution ($\alpha = 0.994$). Di Giovanni et al (2011) perform a similar analysis considering data based on the mandatory reporting of firms' income statements to tax authorities in France.

In general, binning the observations in a sample implies a loss of information with respect to the original sample. Intuitively, the reason is that, after binning, we only know how many observations are included in a certain interval, but not their exact location. More formally,

and focusing on the problem at hand, the (adjusted) frequency used by Axtell (2001) and Di Giovanni et al (2011), located at the geometric mean of the bin endpoints, is not a sufficient statistic for the Pareto shape parameter, and this results in a loss of information. Given the data, as the number of bins gets smaller, the loss of information increases, because all intervals become wider. Hence, any statistical inference procedure based on binned data produces less reliable results than the same procedure based on the actual values of the observations (Virkar and Clauset, 2014). Similar problems arise when only a small sample of the full distribution is available (Perline, 2005; Segarra and Teruel, 2012).

In our analysis we exploit a comprehensive dataset covering the universe of French firms: the data are analogous to those used in Eaton et al (2011) and have been used elsewhere as well (e.g. Garicano et al, 2013). The data on total revenues that we use to measure firm size are taken from the FICUS (*Fichier complet de Système Unifié de Statistique d'Entreprises*) database maintained by the French National Statistical Office (INSEE). We focus on the year 2003 (although the choice of year is actually irrelevant in terms of the results), and have information on more than 2 million firms, excluding the very few cases in which a firm reports total revenues equal to zero. To investigate the relationship between firm scope and size distribution (Section 3 below), we use the information collected by the French Customs, which reports values, destinations and product classes of export flows involving French firms. Our definition of a product is therefore a 6-digit code within the Harmonized System classification. We take the number of different products exported and/or the number of foreign destinations they served as a proxy for firm scope. Unfortunately, no comparable information is available for domestic transactions. Firms exporting less than 1 000 euros outside the EU, or less than 100 000 euros within the EU are not required to report their transactions; other than that, the dataset is comprehensive. Furthermore, we can link the two sources of information by means of a unique firm identification number.

Table 1 reports summary statistics for the full sample under investigation, as well as for a series of subgroups based on the number of products exported and/or destination served

by firms operating in international markets. Exporters represent 4.1% of French firms, but are significantly larger than firms serving only the domestic market. More than one third of firms export just one product, 42.5% of them ship to one destination only, and 29.3% sell one product to one foreign market. On the other hand, wide-scope (highly-diversified) firms represent a small fraction of the universe (0.2%), but they account for more than 25% of total sales.

Table 1: Summary statistics: number of firms and average size (full sample and subgroups)

| | # firms | share | share of exporters | share of total sales | average firm size |
|---|---|---|---|---|---|
| full sample | 2,247,547 | 100% | | 100% | 1,333 |
| exporters | 92,057 | 4.1% | 100% | 59.2% | 19,281 |
| *of which:* | | | | | |
| single product | 31,387 | 1.4% | 34.1% | 5.3% | 5,067 |
| multi-product | 60,670 | 2.7% | 65.9% | 53.9% | 26,604 |
| more than 5 prod. | 28,150 | 1.3% | 30.6% | 45.1% | 47,694 |
| more than 10 prod. | 16,084 | 0.7% | 17.5% | 38.9% | 71,767 |
| single destination | 39,122 | 1.7% | 42.5% | 6.2% | 4,785 |
| multiple destinations | 52,935 | 2.4% | 57.5% | 53.0% | 29,994 |
| more than 5 dest. | 23,651 | 1.1% | 25.7% | 42.9% | 54,297 |
| more than 10 dest. | 13,249 | 0.6% | 14.4% | 36.1% | 81,715 |
| single prod.-dest. | 26,985 | 1.2% | 29.3% | 4.2% | 4,693 |
| multiple prod.-dest. | 65,072 | 2.9% | 70.7% | 55.0% | 25,321 |
| more than 5 prod.-dest. | 36,974 | 1.6% | 40.2% | 48.4% | 39,118 |
| more than 10 prod.-dest. | 25,712 | 1.1% | 27.9% | 44.5% | 51,626 |
| more than 50 prod.-dest. | 7,463 | 0.3% | 8.1% | 32.6% | 129,931 |
| more than 100 prod.-dest. | 3,527 | 0.2% | 3.8% | 25.4% | 213,644 |

Firm size measured in terms of sales (1,000 of euros).

# 3 Empirical Analysis

The empirical analysis starts by investigating the features of the whole firm size distribution by means of the ME approach, which is a method for fitting nonparametric density obtained by maximizing the Shannon's information entropy under constraints that impose the equality of the first $k$ theoretical and empirical moments; see Kapur (1989) for details. Then we run the tests on the original data without sampling or binning observations in size groups. Afterwards

we concentrate on the upper tail of the distribution and investigate the presence of power-law behavior using the three different methodologies. Last, we look at the relationship between the scope of firms and the properties of the size distribution.

Figure 1 displays the histogram of the logarithms of the data along with the truncated normal (logarithm of the truncated lognormal), the exponential (logarithm of Pareto) and the best fitting ME distributions.[1] We use observations larger than 14 000 euros, as below this threshold the distribution is very irregular. Only the smallest 3.7% of the observations are discarded in this way. The optimal ME distribution has $k = 4$; hence, neither the Pareto ($k = 1$) nor the lognormal ($k = 2$) distributions are good approximations, although it is clear from the graph that the lognormal one is "closer" to the true distribution than the Pareto distribution.

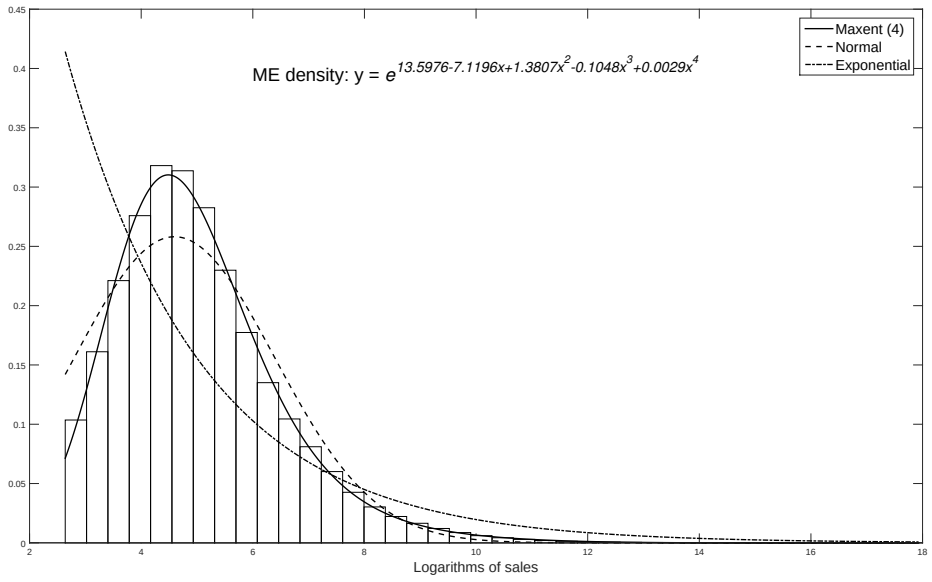


Figure 1: The size distribution of the natural logarithm of French firms, along with the exponential (log of Pareto), normal (log of lognormal) and ME densities. The best-fitting ME density has $k = 4$, and its functional form is given in the plot. Sales are in thousand of euros.

[1]The MLEs of the truncated normal are computed by means of the EM algorithm.

To reinforce the claim that the Pareto distribution does not provide a very good approximation of the data, and to understand why much of the existing literature, at least starting from Axtell (2001), reaches the opposite conclusion (Gabaix, 2009), we investigate the impact of binning the data on the result.

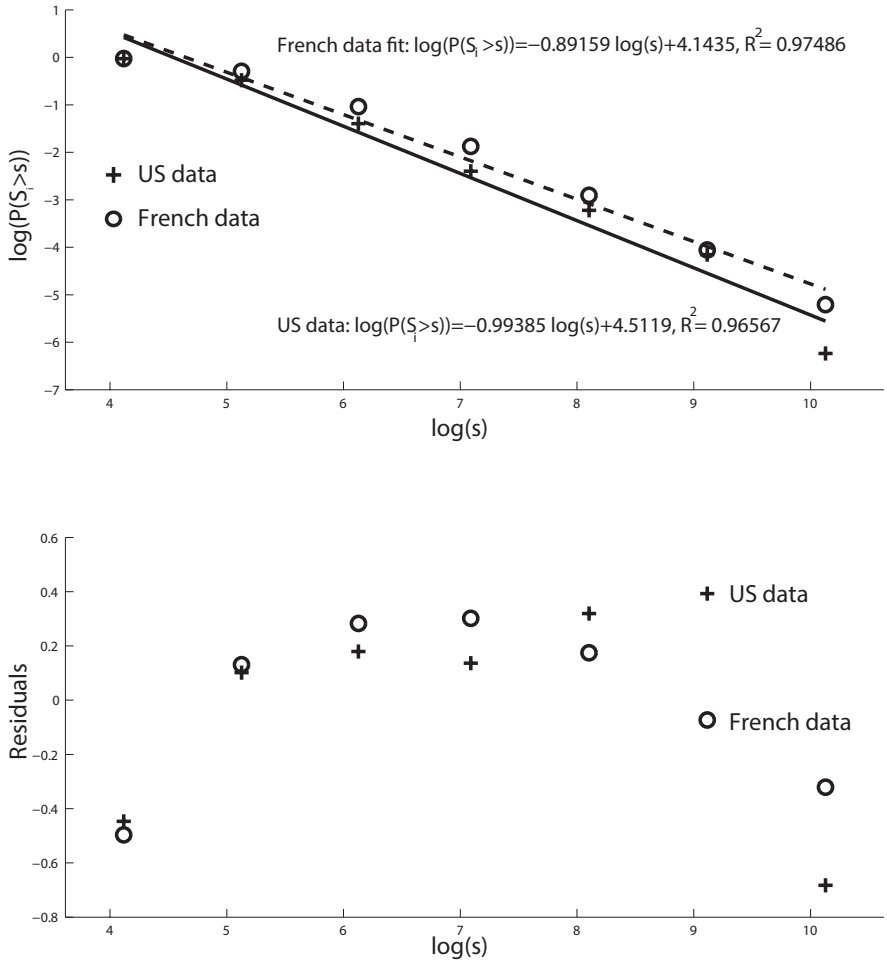


Figure 2: The size distribution of US and French firms: the upper panel shows the complementary cumulative distribution function of bin frequencies in double log scale; the lower panel displays the residuals of the linear regressions on binned data.

The upper panel of Figure 2 shows the distribution of US and French firm sizes. The

graph is the log-log plot of the counter-cumulative distribution function of bin frequencies. For comparison purposes, we use the same data and number of bins (7) used by Axtell (2001), and the bins are equally spaced on a logarithmic scale. The results look qualitatively similar. Nevertheless, the slope of the regression line is considerably smaller in absolute value for French firms. The residuals from the linear regressions are shown in the lower panel of the same figure: they look far from random, showing instead a clear cyclical pattern. These doubts are confirmed by performing the test of power-law behavior with binned data developed by Virkar and Clauset (2014)[2]: at the 5% level, the null hypothesis of power-law is rejected for both the US data ($p$-value = 0.02) and the French data ($p$-value smaller than 0.001).

Furthermore, Virkar and Clauset (2014) note that "for small values of $n$, or for a small number of bins[...], the empirical distribution may closely follow a power-law shape, yielding a large $p$[-value], even if the underlying distribution is not a power law". Thus, whereas the risk of a false positive (the test finds a power-law when the true distribution is not power-law) is high, a false negative does not seem to be a major concern.

Having said that the full distribution is not well fitted neither by a Pareto distribution nor by a lognormal one, we now consider the tail behavior. According to the discussion in Section 2, we should trust the tests that use raw data rather than binned data more. Thus, we carried out the UMPU, ME and GI tests using the 2 247 547 observations on the receipts of French firms. The results are reported in Table 2.[3] The standard errors for $\hat{\alpha}^{UMPU}$ and $\hat{\alpha}^{ME}$ are computed by means of non-parametric bootstrap with 100 replications,[4] whereas

[2]The test is an extension to the case of binned data of the bootstrap-based test for testing the power-law hypothesis developed by Clauset et al (2009). This approach estimates the parameters via maximum likelihood and uses the Kolmogorov-Smirnov goodness-of-fit statistic and the likelihood ratio tests for making a decision.

[3]The ME methodology lends itself in a natural way to testing the hypothesis $H_0 : k = 2$ against $H_1 : k = 3$. As the ME(2) distribution is lognormal, this test allows to find the length of a (possible) lognormal tail. At the 5% level, the test finds a lognormal tail containing the largest 47 000 firms, approximately corresponding to the 0.021 quantile and to the 75.59% of the total revenues.

[4]The standard error of the Hill estimator $\hat{\alpha}^{UMPU}$ can in principle be computed analytically using the asymptotic normality of the estimator. However, the regularity conditions necessary for this result are difficult to verify (Embrechts et al, 1997, p. 337-339), so that the bootstrap estimate is more reliable.

Table 2: Test results for French firms ($n = 2\,247\,547$): power-law threshold rank, power-law threshold, percentile, share and shape parameter. The power-law threshold is the observation corresponding to the rank found by each test.

| | UMPU | | ME | | GI | |
|---|---|---|---|---|---|---|
| | 5% level | 1% level | 5% level | 1% level | 5% level | 1% level |
| rank | 1600 | 1650 | 1750 | 2150 | 2400 | 3480 |
| power-law threshold | 179 337 | 173 321 | 162 246 | 134 022 | 119 177 | 84 483 |
| percentile | $< 0.1$ | $< 0.1$ | $< 0.1$ | 0.1 | 0.1 | 0.2 |
| % revenues | 41.89 | 42.19 | 42.75 | 44.70 | 45.75 | 49.35 |
| shape | 1.158 | 1.147 | 1.137 | 1.111 | 1.181 | 1.151 |
| s.e. | 0.062 | 0.061 | 0.047 | 0.042 | 0.020 | 0.016 |
| 95% CI | [1.101,1.352] | [1.096,1.403] | [1.062,1.254] | [1.031,1.273] | [1.141,1.221] | [1.110,1.192] |

The values in the table should be interpreted as follows: *rank* gives the number of observations larger than the power-law threshold, *percentile* is the percentile of the empirical distribution corresponding to the rank, *% revenues* is the percentage of total revenues corresponding to the firms in the power-law tail, *shape* is the estimate of the Pareto shape parameter, s.e. is the standard error, 95% CI is the 95% confidence interval.

the asymptotic standard error of $\hat{\alpha}^{GI}$ is given by $(\hat{\alpha}^{GI})^2/\sqrt{2n}$ (Gabaix and Ibragimov, 2011). Similarly, the confidence interval for $\hat{\alpha}^{UMPU}$ and $\hat{\alpha}^{ME}$ is the empirical confidence interval of the corresponding bootstrap distribution, whereas for $\hat{\alpha}^{GI}$ it is the confidence interval of the limiting normal distribution. Although the GI test finds a slightly longer tail than UMPU and ME, the overall messages are that the Pareto tail (if any) corresponds to a quite small fraction of the largest firms and the shape parameter is significantly different from 1.

Even though these very few observations contain a non-negligible share of the total receipts, to confirm the lack of statistical significance of the Pareto tail found by the tests we carry out a simulation-based analysis. The description of the experiment is as follows. We sample $n$ observations from a lognormal distribution with parameters estimated from the real data and estimate the probability $\hat{p}$ of obtaining a power-law tail at least as long as the one found by the tests in the real data:

$$\hat{p} = \frac{\#\{r_i > r^*\}}{B},$$

where $B = 500$ is the number of replications, $r^*$ is the rank obtained applying the test to

the real data, $r_i$ $(i = 1, \ldots, B)$ is the rank such that, at the $i$-th replication, the test is below the 95% critical value for ranks $1, \ldots, r_i - 1$ and above this value for ranks $r_i, r_i + 1, \ldots, n$. A small value of $\hat{p}$ implies that the data display a tail which is significantly longer than the tail found for the lognormal used in the simulation.

We obtain $\hat{p} = 0.71$ for ME, $\hat{p} = 0.62$ for UMPU and $\hat{p} = 0.25$ for GI. These outcomes suggest that the Pareto tail found by the tests in the real data is not significantly longer than the one identified when applying the tests to a lognormal distribution. In sum, according to these outcomes, the size distribution of French firms is not Pareto, even in the tail.

It is worth extending the study of the tail behavior also to the subpopulations obtained when one moves to consider firm diversification and internationalization levels. This is done by discriminating among different subsets of firms, based on the number of products they sell, the number of foreign destinations they serve, and the number of product-destination pairs.

Table 3 reports the test results for various subsets of firms. The first row of panel *a* is based on the whole dataset and thus simply replicates Table 2, the second ($K > 0$) refers to exporting firms only, while the next three lines concern multi-product firms. The number of observations shrinks significantly when we move from the universe of firms in the dataset to exporting firms only (92 000 observations, roughly 4% of the total), whereas only 16 000 firms export more than 10 different products. Although the power of the tests decreases with the sample size, most of these numbers are still sufficiently large to guarantee reliable results. Moreover, when considering the smallest subsets, we are mostly interested in the point estimate of the shape parameter, whose standard error incorporates the effect of the sample size. All three tests find a monotonic increase in the share of firms belonging to the Pareto tail of the distribution, whose length goes from virtually nil (0.07–0.15% depending on the test) to a value ranging between 5 and 8.7% of the distribution. Similarly, the estimated shape parameter of the power-law decreases monotonically toward 1 once we progressively restrict the analysis to firms exporting a larger number of products.

Table 3: Test results for firms with different levels of diversification (number of products, $K$) and internationalization (number of foreign markets, $N$, and product-market pairs, $NK$)

| | UMPU | | | ME | | | GI | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | panel *a*: number of products | | | | | | |
| $K$ | rank | perc. | shape | rank | perc. | shape | rank | perc. | shape | # firms |
| $\geq 0$ | 1600 | $< 0.1$ | 1.158 (0.062) [1.101,1.352] | 1750 | $< 0.1$ | 1.137 (0.047) [1.062,1.254] | 2400 | 0.1 | 1.181 (0.020) [1.141,1.221] | 2 247 547 |
| $> 0$ | 1350 | 1.5 | 1.136 (0.041) [1.076,1.255] | 1450 | 1.6 | 1.112 (0.046) [1.044,1.234] | 2410 | 2.6 | 1.135 (0.019) [1.096,1.174] | 92 057 |
| $> 1$ | 1300 | 2.1 | 1.123 (0.056) [1.045,1.269] | 1350 | 2.2 | 1.112 (0.063) [0.991,1.270] | 2230 | 3.7 | 1.125 (0.019) [1.084,1.166] | 60 670 |
| $> 5$ | 1050 | 3.7 | 1.089 (0.051) [0.992,1.210] | 1150 | 4.1 | 1.057 (0.048) [0.972,1.150] | 1830 | 6.5 | 1.086 (0.020) [1.041,1.131] | 28 150 |
| $> 10$ | 850 | 5.3 | 1.073 (0.062) [0.966,1.213] | 950 | 5.9 | 1.049 (0.065) [0.929,1.143] | 1400 | 8.7 | 1.072 (0.022) [1.020,1.124] | 16 084 |
| | | | | panel *b*: number of destinations | | | | | | |
| $N$ | rank | perc. | shape | rank | perc. | shape | rank | perc. | shape | # firms |
| $> 1$ | 1250 | 2.4 | 1.124 (0.050) [1.038,1.233] | 1350 | 2.6 | 1.106 (0.053) [0.997,1.187] | 2235 | 4.2 | 1.118 (0.019) [1.077,1.159] | 52 935 |
| $> 5$ | 1050 | 4.8 | 1.113 (0.054) [0.998,1.194] | 1100 | 4.6 | 1.043 (0.059) [0.947,1.171] | 1810 | 7.6 | 1.070 (0.019) [1.025,1.115] | 23 651 |
| $> 10$ | 850 | 6.4 | 1.109 (0.050) [0.937,1.142] | 900 | 6.8 | 1.018 (0.054) [0.888,1.095] | 1460 | 11.0 | 1.040 (0.020) [0.989,1.091] | 13 249 |
| | | | | panel *c*: number of product-destination pairs | | | | | | |
| $NK$ | rank | perc. | shape | rank | perc. | shape | rank | perc. | shape | # firms |
| $> 1$ | 1300 | 2.0 | 1.133 (0.050) [1.037,1.232] | 1350 | 2.1 | 1.122 (0.055) [1.002,1.212] | 2280 | 3.5 | 1.127 (0.019) [1.086,1.167] | 65 072 |
| $> 5$ | 1150 | 3.1 | 1.136 (0.035) [1.030,1.150] | 1200 | 3.2 | 1.081 (0.044) [0.980,1.139] | 2095 | 5.7 | 1.098 (0.019) [1.056,1.140] | 36 974 |
| $> 10$ | 1050 | 4.1 | 1.119 (0.048) [1.024,1.227] | 1150 | 4.5 | 1.082 (0.046) [0.976,1.186] | 1800 | 7.0 | 1.088 (0.020) [1.042,1.134] | 25 712 |
| $> 50$ | 780 | 10.4 | 1.082 (0.055) [0.937,1.149] | 820 | 11.0 | 0.995 (0.060) [0.872,1.114] | 1195 | 16.0 | 1.031 (0.022) [0.975,1.087] | 7463 |
| $> 100$ | 520 | 14.7 | 1.121 (0.067) [0.857,1.129] | 600 | 17.0 | 0.941 (0.070) [0.777,1.066] | 780 | 22.1 | 1.002 (0.025) [0.933,1.071] | 3527 |

The values in the table should be interpreted as follows: *rank* gives the number of observations larger than the Pareto threshold, *perc.* is the percentile of the empirical distribution corresponding to the rank, *shape* is the estimate of the Pareto shape parameter and the numbers in brackets are the corresponding 95% confidence intervals.

Similar conclusions hold when we look at "very international" firms, i.e. companies shipping their goods to many foreign destinations (see panel *b* of Table 3). Indeed, when compared to the total population or the universe of exporters, the size distribution of firms exporting to more than 10 destinations (which make up less than 1% of all firms) displays a power-law tail spanning between 6 and 13% of the population. Furthermore, the estimated shape parameter moves downward becoming closer to 1.

The change in the behavior of the distribution is all the more apparent when we classify firms on the basis of the number of their product-destination pairs, thus distinguishing between, say, apples shipped to country A and to country B (see panel *c*).

The convergence towards a power-law appears clearly in Table 3 as well. According to the 95% confidence intervals, the hypothesis $\alpha = 1$ is often not rejected for large values of $K$, $N$ and $NK$: in particular, all the confidence intervals of all the tests contain the value 1 when $N > 10$, $NK > 50$ and $NK > 100$, and two tests out of three contain the value 1 when $K > 5$, $K > 10$ and $N > 5$.

The tendency toward Zipf's law in the upper tails of the distributions appears clearly in Figure 3 as well, where we portray the counter-cumulative distribution functions pertaining to the different groups of firms, along with a reference line with a slope equal to $-1$.[5] Interestingly, the top panel shows that the distribution of all firms is similar to the one found by Rossi-Hansberg and Wright (2007) on US firms, with the central part approximately resembling a Pareto distribution and a more pronounced concave shape in the tails. The plots confirm the results presented in Table 3: the Zipf behavior is more apparent in the case of multiple products (top panel) than in the case of firms serving multiple markets. The overall distribution of firm size is a weighted average of the distributions of different subgroups, with weights given by their relative importance in the total population. As a result, the size distributions of business firms may differ depending on the weights of each type. A larger share of diversified firms would lead to a more pronounced power-law behavior and a closer

[5]The various distributions have been right-shifted to improve readability.

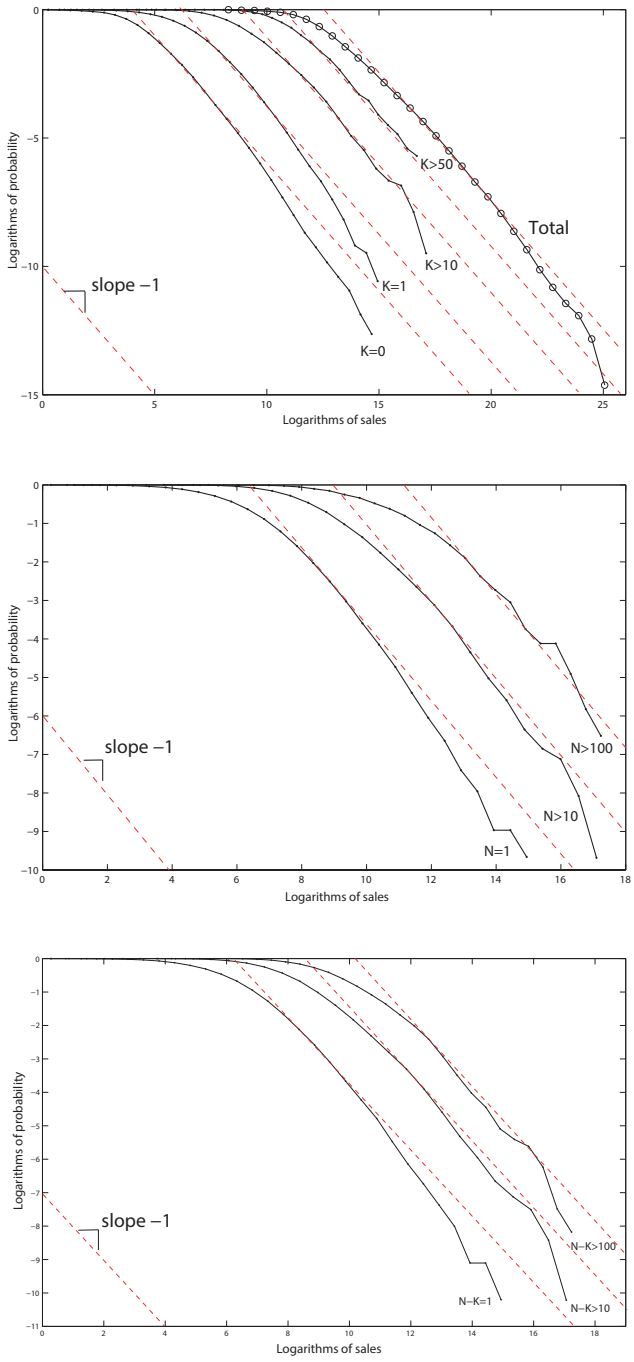


Figure 3: The size distribution of French firms by number of products, $K$, destination markets, $N$ and product-destination pairs $NK$. The distributions have been right-shifted to improve readability.

resemblance to Zipf's law (at least in the upper tail). If, for instance, US firms were generally more diversified and/or more internationalized than French ones, then the corresponding size distribution would display a fatter Pareto tail.

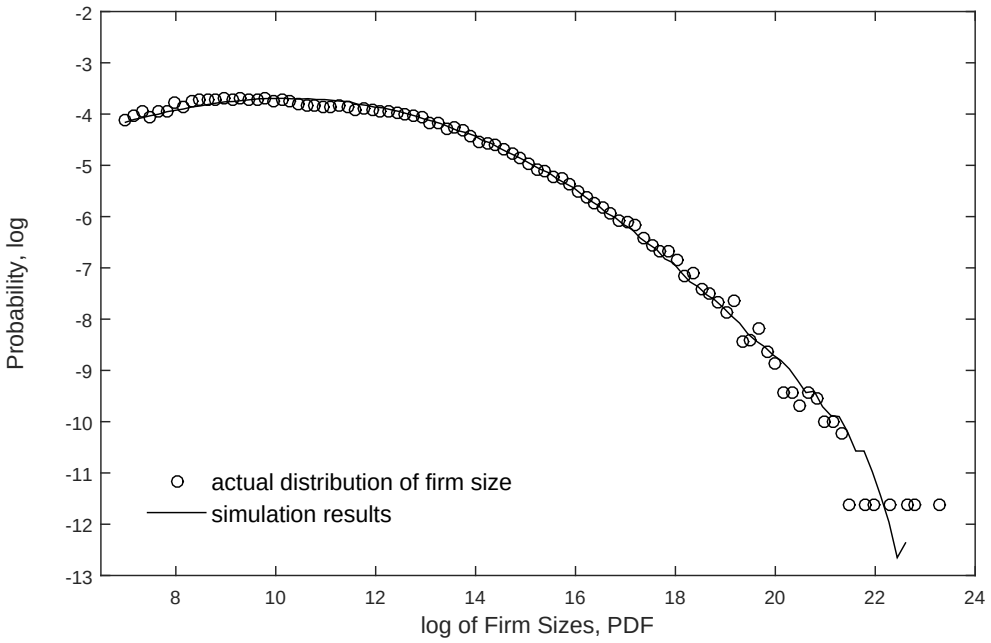


Figure 4: The size distribution of French firms and simulation results.

To better understand the emergence of a Pareto tail for large diversified companies, we follow Growiec et al (2008) and perform a series of simulations; see also (De Fabritiis et al, 2003; Fu et al, 2005; Yamasaki et al, 2006; Buldyrev et al, 2007; Pammolli et al, 2007; Riccaboni et al, 2008). According to their model, the number of products in a firm grows in proportion to the existing number of products and the size of each product grows in proportion to its size. As a result, the Gibrat-like growth process at the level of single products/markets leads to a lognormal distribution of product sizes, whereas the proportionate growth process in the number of products generates a Pareto distribution of the number of products by firm, like in the Simon model (Simon, 1955). Therefore the size distribution of single-product firms is lognormally distributed, as in the original Gibrat model, but the size distribution of

multi-product firms is difficult to determine since the sum of lognormally distributed variates does not have a closed form solution. Our simulations reveal that, as predicted by Growiec et al (2008), the emergence of a Pareto upper tail depends on the interplay between the two growth processes of the number and size of products. We proceed as follows. First, we estimate the parameters of the lognormal distribution of product sizes in France. Next, we generate a Zipf distribution of the number of products to match real world data. Finally, we randomly assign to each firm a number of products drawn from the Zipf's law, and to each product a size drawn from the lognormal distribution. Summing over products for each firm we then compute firm size.

As shown in Figure 4 the simulated distribution of firm sizes closely matches the size distribution of French firm sizes[6]. This is a striking result given that in our simulations we just use the empirically observed distribution of product sizes with no free parameters. Thus we can conclude that the size distribution of firms lies between a lognormal for small, single-product firms and a Pareto distribution which might emerge in the upper tail of large diversified companies. Figure 5 shows what happens when we sum lognormally distributed product sizes. Our simulations run in two steps. First, we generate the number of products $K_i$ of a given firm $i$ by sampling a discrete power law distribution of product numbers. Second, we compute the total sales of firm $i$ as the sum of $K_i$ products whose size is sampled from the lognormal distribution of product sizes (LDPS). We do the same computation for every firm in the industry. Therefore, the size of business firms depends on both the extensive margin (i.e. the number of products sold) and the intensive margin (i.e. the sales of each and every product in its portfolio). When the scale parameter of the size distribution of products is small (i.e. products are approximately of the same size) and the number of products by firm is very heterogeneous (i.e. power-law), the size of firms is Pareto all the way down to small companies (a straight line in double log scale in Figure 5(b)). Conversely, when firms have a

[6]Simulation results are based on 100 replications of the random assignment process. Only firms with more than 1,000 euros of revenue are considered.

small (and similar) number of products of different sizes we observe a lognormal distribution of firm sizes[7]. The actual distribution of firm sizes has a lognormal body and might depict a Pareto upper tail, as postulated by Growiec et al (2008), depending on the interplay between the lognormal distribution of product sizes and the Pareto distribution of product numbers.

# 4 Discussion and conclusions

This paper studies the size distribution of business firms using comprehensive data on French companies. We adopt a rigorous methodological approach that applies both parametric and non-parametric procedures and addresses the shortcomings that characterize many existing studies. We look at the entire distribution of firms investigating the tail behavior of the distribution for diversified and international firms. It is worth noting that all the tests and simulations employed in this paper to identify the tail behavior of the size distributions give approximately the same outcomes: this suggests that our conclusions have a strong empirical justification.

We do not find support for the hypothesis that the size distribution follows either a Zipf's law or a Pareto distribution. This is consistent with recent evidence put forward by, for instance, Rossi-Hansberg and Wright (2007) and Head et al (2014). Even if we consider the power-law as a tail property, the shape parameter is significantly larger than 1, and only approximately 0.1% of the firms belong to the Pareto tail, corresponding to less than 50% of combined total revenues. However, we document a tendency toward the emergence of a Zipf's law upper tail for the group of large multi-product firms, that are more actively engaged in export markets. Based on a simulation exercise, we show how the size distribution of firms can vary from lognormal to Pareto, depending on the interplay of two growth processes in the number and size of products. Our work contributes to shed new light on the role of diversification patterns in the process of firm growth. With few notable

[7]More precisely, the resulting distribution is given by a sum of lognormal distributions that, when $\sigma$ is small, can be approximated by a lognormal distribution.

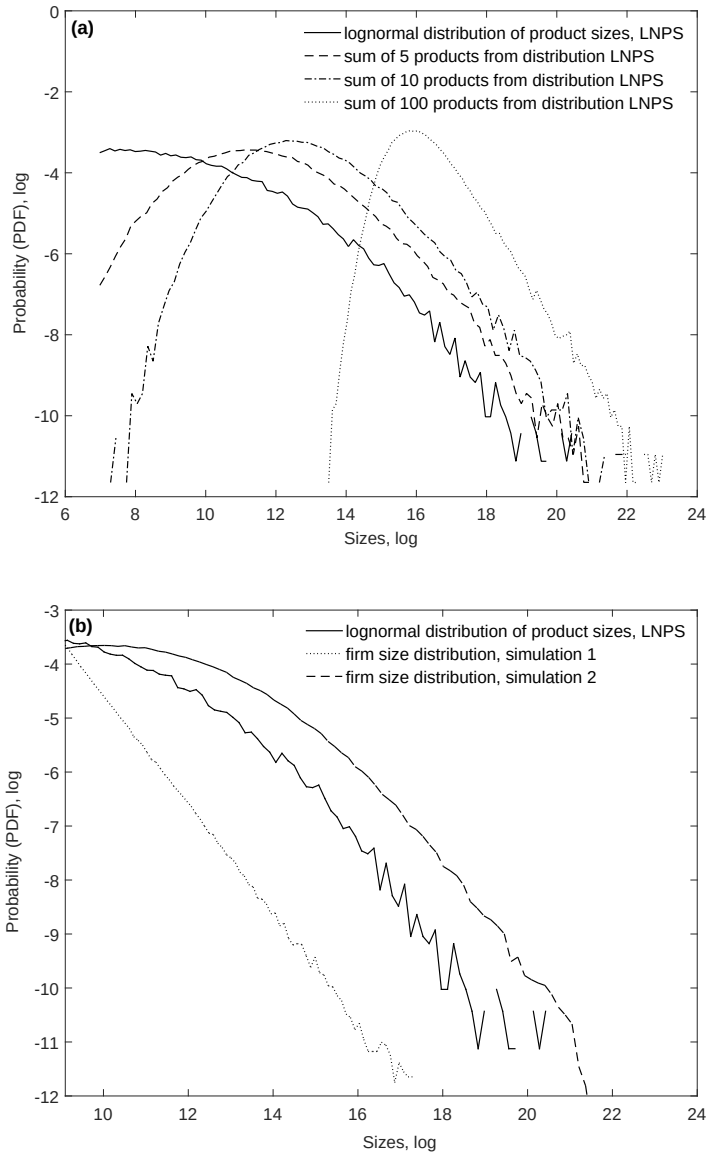


Figure 5: Simulation results. Subplot (a): The lognormal distribution of product sizes (LDPS) and the sum of lognormals. The value of $\mu$ and $\sigma^2$ of the LDPS are the maximum likelihood values of the size distribution of French products: $\mu = 7.613$, $\sigma^2 = 2.833$. We also report the distribution of the sum of 5, 10 and 100 values randomly sampled from the LDPS. Subplot (b): The size distribution of firms is derived by sampling a power-law distributed number of products from the LDPS. In the first simulation we use $\sigma^2/10 = 0.283$ to show that when the variance of the product distribution is small, the power-law behavior of the distribution of the number of products dominates the final firm size distribution. In simulation 2 we use the actual value of $\sigma^2 = 2.833$ for the LDPS and a Pareto distribution of product numbers to obtain the size distribution of firms which depicts a power-law upper tail.

exceptions, the role of firm diversification has been neglected in that literature. In more general terms, our paper contributes to the recent body of work that explains departures from benchmark distributions, either Pareto or Lognormal, and differences across sectors or firm types. Our results are compatible with multiple theoretical models recently appeared in the literature. For instance, Chatterjee and Rossi-Hansberg (2012) show that so long as new ideas are captured mostly by established firms (versus startups), the size distribution of firms converges to Zipf's law. Since few firms are highly diversified from the very beginning, and few new exporters ship many goods to many destinations the first time they enter foreign markets (Albornoz et al, 2012), when looking at firms with wide scope, we are focusing on established ventures. Similarly, Rossi-Hansberg and Wright (2007) argue that higher human capital intensity leads to an approximate power-law size distribution. Since the empirical literature on firm behavior in international markets has found a link between the degree of international involvement and human capital at the firm level (Munch and Skaksen, 2008), our empirical findings are consistent with this interpretation as well.

From a practical point of view we consider the lognormal distribution to be a good approximation of the size distribution of French firms; departures from this benchmark in the lower and upper tail are informative of economic forces and frictions, and deserve further scrutiny. For instance, lognormality allows for the easy derivation of concentration indexes and therefore for the quantification of departures from a clear-cut theoretical benchmark (on this subject see Hart, 1975; Davies, 1980). Overall, the actual size distribution of firms largely depends on the degree of diversification of economic activities across (almost independent) products and markets. When firms are diversified across many independent areas of activities we tend to observe the emergence of a Pareto upper tail in the distribution of firm sizes. Therefore, it makes much sense to devote future research to the analysis of differences across countries and industries, and to cross-compare diversification patterns of firms in different empirical domains.

# References


Albornoz F, Calvo Pardo HF, Corcos G, Ornelas E (2012) Sequential exporting. Journal of International Economics 88(1):17–31

Axtell RL (2001) Zipf distribution of U.S. firm sizes. Science 293(5536):1818–1820

Bee M, Riccaboni M, Schiavo S (2011) Pareto versus lognormal: A maximum entropy test. Physical Review E 84:026,104, DOI 10.1103/PhysRevE.84.026104

Bee M, Riccaboni M, Schiavo S (2013) The size distribution of US cities: Not Pareto, even in the tail. Economics Letters 120(2):232–237

Buldyrev SV, Growiec J, Pammolli F, Riccaboni M, Stanley HE (2007) The growth of business firms: Facts and theory. Journal of the European Economic Association 5(2-3):574–584

Cabral LMB, Mata J (2003) On the evolution of the firm size distribution: Facts and theory. American Economic Review 93(4):1075–1090

Chatterjee S, Rossi-Hansberg E (2012) Spinoffs and the market for ideas. International Economic Review 53(1):53–93

Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Review 51:661–673

Coad A (2009) The Growth of Firms: A Survey of Theories and Empirical Evidence. Edward Elgar, Cheltenham

Davies S (1980) Measuring industrial concentration: An alternative approach. Review of Economics and Statistics 62(2):306–309

De Fabritiis G, Pammolli F, Riccaboni M (2003) On size and growth of business firms. Physica A: Statistical Mechanics and its Applications 324(1-2):38–44

Del Castillo J, Puig P (1999) The best test of exponentiality against singly truncated normal alternatives. Journal of the American Statistical Association 94:529–532

Di Giovanni J, Levchenko AA, Rancière R (2011) Power laws in firm size and openness to trade: Measurement and implications. Journal of International Economics 85(1):42–52, DOI DOI: 10.1016/j.jinteco.2011.05.003

Eaton J, Kortum S, Kramarz F (2011) An anatomy of international trade: Evidence from French firms. Econometrica 79(5):1453–1498

Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling Extremal Events for Insurance and Finance. Springer

Fu D, Pammolli F, Buldyrev SV, Riccaboni M, Matia K, Yamasaki K, Stanley HE (2005) The growth of business firms: Theoretical framework and empirical evidence. Proceedings of the National Academy of Sciences of the United States of America 102(52):18,801–18,806, DOI 10.1073/pnas.0509543102

Gabaix X (2009) Power laws in economics and finance. Annual Review of Economics 1:255–93

Gabaix X, Ibragimov R (2011) Rank-1/2: A simple way to improve the OLS estimation of tail exponents. Journal of Business and Economic Statistics 29(1):24–39

Garicano L, LeLarge C, Van Reenen J (2013) Firm size distortions and the productivity distribution: Evidence from france. Working Paper 18841, NBER

Gibrat R (1931) Les Inegalites Economiques. Sirey, Paris

Growiec J, Pammolli F, Riccaboni M, Stanley HE (2008) On the size distribution of business firms. Economics Letters 98(2):207–212

Halvarsson D (2013) Industry Differences in the Firm Size Distribution. Ratio Working Papers 214, The Ratio Institute

Hart PE (1975) Moment distributions in economics: An exposition. Journal of the Royal Statistical Society Series A (General) 138(3):423–434

Hart PE, Prais SJ (1956) The analysis of business concentration: A statistical approach. Journal of the Royal Statistical Society Series A (General) 119(2):150–191

Head K, Mayer T, Thoenig M (2014) Welfare and trade without Pareto. The American Economic Review 104(5):310–316

Huber P, Pfaffermayr M (2010) On measuring business concentration. Oxford Bulletin of Economics and Statistics 72:648–668

Ijiri Y, Simon HA (1974) Interpretations of departures from the Pareto curve firm-size distributions. Journal of Political Economy 82(2):315–331

Jovanovic B (1982) Selection and the Evolution of Industry. Econometrica 50(3):649–70

Kapur J (1989) Maximum entropy models in science and engineering. Wiley

Klette TJ, Kortum S (2004) Innovating firms and aggregate innovation. Journal of Political Economy 112(5):986–1018

Kumar KB, Rajan RG, Zingales L (1999) What determines firm size? Working Paper 7208, NBER

Luttmer EG (2007) Selection, growth, and the size distribution of firms. Quarterly Journal of Economics 122(3):1103–1144

Luttmer EGJ (2011) On the Mechanics of Firm Growth. Review of Economic Studies 78(3):1042–1068

Malevergne Y, Pisarenko V, Sornette D (2009) Gibrat's law for cities: uniformly most powerful unbiased test of the Pareto against the lognormal. Swiss Finance Institute Research Paper Series pp 09–40

Malevergne Y, Pisarenko V, Sornette D (2011) Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. Physical Review E 83(3):036,111, DOI 10.1103/PhysRevE.83.036111

Melitz MJ (2003) The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica 71(6):1695–1725

Mitzenmacher M (2004) A brief history of generative models for power law and lognormal distributions. Internet mathematics 1(2):226–251

Munch JR, Skaksen JR (2008) Human capital and wages in exporting firms. Journal of International Economics 75(2):363–372

Pammolli F, Fu D, Buldyrev SV, Riccaboni M, Matia K, Yamasaki K, Stanley HE, et al (2007) A generalized preferential attachment model for business firms growth rates. European Physical Journal B Condensed Matter Physics 57(2):127

Perline R (2005) Weak and false inverse power laws. Statistical Science 20:68–88

Riccaboni M, Pammolli F, Buldyrev SV, Ponta L, Stanley HE (2008) The size variance relationship of business firm growth rates. Proceedings of the National Academy of Sciences 105(50):19,595–19,600, DOI 10.1073/pnas.0810478105

Rossi-Hansberg E, Wright ML (2007) Establishment size dynamics in the aggregate economy. American Economic Review 97(5):1639–1666

Segarra A, Teruel M (2012) An appraisal of firm size distribution: Does sample size matter? Journal of Economic Behavior & Organization 82(1):314–328

Simon HA (1955) On a class of skew distribution functions. Biometrika 42(3-4):425

Simon HA, Bonini CP (1958) The size distribution of business firms. American Economic Review 48(4):607–617

Sutton J (1997) Gibrat's legacy. Journal of Economic Literature 35(1):40–59

Tang A (2015) Does Gibrat's law hold for Swedish energy firms? Empirical Economics 49:659–674

Virkar Y, Clauset A (2014) Power-law distributions in binned empirical data. Annals of Applied Statistics 8(1):89–119

de Wit G (2005) Firm size distributions: An overview of steady-state distributions resulting from firm dynamics models. International Journal of Industrial Organization 23(5-6):423–450

Yamasaki K, Matia K, Buldyrev SV, Fu D, Pammolli F, Riccaboni M, Stanley HE (2006) Preferential attachment and growth dynamics in complex systems. Physical Review E 74(3):035,103