

The size distribution across all cities - double Pareto lognormal strikes!☆

Kristian Giesen, Arndt Zimmermann, Jens Suedekum*

Mercator School of Management, University of Duisburg-Essen

Abstract

Using un-truncated settlement size data from eight countries, we show that the "double Pareto lognormal" (DPLN) distribution provides a better fit to actual city sizes than the simple lognormal (LN) distribution. The DPLN has a lognormal body and features a power law in both the lower and the upper tail. It emerges in the steady-state of a stochastic urban growth process with random city formation. Our findings reconcile a recent debate on the Zipfian rank-size rule for city sizes.

Key words: Zipf's law, urban growth, Gibrat's law, city size distributions

JEL: R11, R12, O4

1. Introduction

Recently there has been an intensive debate about city size distributions. Dozens of older studies have argued that city sizes follow a Pareto distribution, or even adhere exactly to the well known rank-size rule known as Zipf's law.¹ This evidence is problematic, however, since those studies have worked with truncated samples and focussed only on large cities. In an influential article, Eeckhout (2004) has shown that the Pareto does *not* hold when taking into account *all* settlements of a country. Figures 1a-1c below illustrate this point. The solid lines depict, respectively, the entire city size distribution in Germany, the United States, and France (in logarithmic scale). It is immediately obvious that these are no Pareto distributions, i.e., that Zipf's law does not hold across all cities in these countries. This raises three important questions. First, if not the Pareto, what is the appropriate parameterization for city sizes? Second, what can we learn from city size distributions about the underlying urban growth process? And third, has Eeckhout (2004) invalidated the entire old Zipf literature, or is there a way to reconcile them?

☆We thank Stuart Rosenthal (the editor), two anonymous referees, and seminar participants at the 2009 North American Regional Science Council in San Francisco. We are particularly grateful to Bill Reed, Jan Eeckhout and Moshe Levy for very helpful comments on an earlier draft of this paper. All errors are solely our responsibility.

*corresponding author: Jens.Suedekum@uni-due.de. Mercator School of Management, University of Duisburg-Essen, Lotharstrasse 65, 47057 Duisburg, Germany

¹Zipf's law states that city sizes are Pareto-distributed with shape parameter equal to minus one. This implies that city sizes follow a particular power law such that the country's largest city is twice as large as the second-largest, three times as large as the third-largest city, and so on. See Soo (2005) and Nitsch (2005) for comprehensive analyses.

Eeckhout (2004) addresses all three questions. He develops a model where cities grow stochastically, and this growth process - the pure form of Gibrat's law - generates a lognormal (LN) size distribution. Eeckhout then shows that the LN delivers a good fit to actual city sizes in the US. This may answer the first two, but has delicate implications for the third question. As a matter of fact, the LN does not feature a power law in the upper tail and, hence, it is strictly speaking not compatible with Pareto and Zipf. Why have so many previous studies, including the recent one by Levy (2009) who uses an un-truncated sample, then provided evidence for a Zipfian power law among large cities? The reason according to Eeckhout (2004, 2009) is that the LN and the Pareto distribution have similar properties in the upper tail and can become virtually indistinguishable. In other words, his answer to the third question is that Zipf can be observed among large cities in practice, because the Pareto closely resembles the true size distribution (the LN) in the top range.²

FIGURES 1a - 1c HERE

In this paper we take a fresh look at these issues by using un-truncated settlement size data from eight countries. We confirm that the city size distribution in all countries can indeed be well approximated by a LN. However, that does not mean that there can be no other distribution which is even more successful in fitting the data. In fact, using various methods we show that the "double Pareto lognormal" (DPLN) distribution consistently provides an even closer fit. That distribution has a lognormal body in the medium range and exhibits a power law in both the lower and the upper tail. Taken by itself it is not surprising that one can find a distribution with a more flexible functional form that delivers a better fit, but we also show that the DPLN is the preferred model according to several statistical selection criteria that penalize it for having more parameters than the LN. Furthermore, the important difference between the DPLN and an arbitrary flexible distribution is that it has a theoretical foundation in terms of the underlying urban growth process. Reed (2002) has shown that the DPLN emerges in the steady-state of an evolutionary process which can be thought of as a generalization of Gibrat's law. Put differently, it is not the intention of this paper to fit just any arbitrary distribution in a theory-free manner. We rather provide an alternative answer to the first two questions by pointing at an urban growth theory (developed in Reed 2002) which generates a size distribution that is even closer to the actual data than the LN.

Finally, our findings may be particularly useful because we can provide a more satisfactory answer to the third question and reconcile the recent debate about city size distributions. In common with Eeckhout (2004, 2009) we find that the LN does a good job in fitting the data. Yet, the empirical distributions in all countries exhibit a distinctive power law pattern in the tails, as also noted by Levy (2009) for the large cities in the US. Though the LN can be consistent with such a power law pattern under certain conditions (see Mitzenmacher 2003), this feature of the data is more precisely captured by the DPLN. In other words, the DPLN is *even better* compatible with Zipf's law among large cities while following a lognormal shape in other ranges. Our results may therefore bring Eeckhout and the older Zipf literature even closer together. In contrast to that literature, which has mostly drawn conclusions from truncated samples, in our case the Zipfian power law emerges as an upper tail feature of the un-truncated distribution.

²Also see Mitzenmacher (2003), who shows that the density or the countercumulative distribution function of the LN generate a "nearly straight" line in logarithmic plots when the variance is large. A power law (Pareto) would generate exactly a straight line in such plots. For the Zipfian rank-size rule to hold, such a straight line is required.

2. Urban growth processes and steady-state city size distributions

In the model by Eeckhout (2004) an economy consists of a fixed number of locations across which workers are freely mobile. The spatial equilibrium results from a trade-off between positive and negative size externalities that accrue within but do not spill over across locations. In every time period, each location is hit by an idiosyncratic and random productivity shock. Cities eventually grow according to the pure form of Gibrat's law, which can be described as $dPop_{it}/Pop_{it} = \mu dt + \sigma dB_{it}$, where $dPop_{it}/Pop_{it}$ is the percentage change of population in city i at time t . The parameter μ is trend growth, and B_{it} is an independent shock with mean zero and variance σ^2 . As already anticipated by Gibrat (1931), such a stochastic proportionate growth process with additive random shocks asymptotically leads to a lognormal (LN) distribution.³

Reed (2002) develops a model that is more statistical in nature. Cities grow stochastically as under Gibrat's law, but in every time interval dt there is the probability λdt that a new city emerges as a satellite of an existing one.⁴ The initial size of the new city is drawn from a LN distribution with mean μ_0 and variance σ_0^2 . These new cities then also exhibit proportionate growth. At time t there are $e^{\lambda t}$ cities in total, some of which are older than others. Reed (2002) proves that this growth process, which resembles the Yule-process first described in biology (see Yule 1925), asymptotically leads to a "double Pareto lognormal" (DPLN) distribution, with density

$$f(x) = \frac{\alpha\beta}{\alpha + \beta} \left[x^{-\alpha-1} e^{\left(\alpha\mu_0 + \frac{\alpha^2\sigma_0^2}{2}\right)} \Phi\left(\frac{\log(x) - \mu_0 - \alpha\sigma_0^2}{\sigma_0}\right) + x^{\beta-1} e^{\left(-\beta\mu_0 + \frac{\beta^2\sigma_0^2}{2}\right)} \Phi^c\left(\frac{\log(x) - \mu_0 + \beta\sigma_0^2}{\sigma_0}\right) \right].$$

α and β are the Pareto coefficients for the upper and the lower tail, respectively, and μ_0 and σ_0 are the lognormal body parameters. Φ represents the normal cumulative density function (cdf) and $\Phi^c = 1 - \Phi$ represents the complementary cdf.

Details about the properties of this distribution can be found in Reed and Jorgensen (2005). It is shown there, for example, that a DPLN distributed random variable X can be represented as UV_1/V_2 , where U , V_1 and V_2 are independent and U is a LN distribution with parameters μ_0 and σ_0 and V_1 and V_2 are Pareto distributions with shape parameters α and β , respectively. The DPLN is unimodal if $\beta > 1$ and can be written as a mixture of a right-handed and a left-handed Pareto-lognormal limiting distribution which, respectively, arises if $\alpha \rightarrow \infty$ or $\beta \rightarrow \infty$. It is not possible to exactly delineate the lognormal body part and the Pareto-distributed tails. That is, we cannot pin down parametrically at which city size the upper tail of the DPLN starts (or where the lower tail ends), although informal approximations are of course possible. Last, the simple LN distribution is nested in the DPLN if $\alpha \rightarrow \infty$ and $\beta \rightarrow \infty$.

Our paper can be seen as the first attempt in the literature to discriminate between the two theories of urban growth, the pure Gibrat's law and the generalized version by Reed (2002). We do so by comparing which of the theoretical steady-state distributions, LN or DPLN, is the preferred model for empirical city size distributions. So far, the fit of these distributions has only been addressed separately. For the DPLN this has been done by Reed (2002), but only for four regions (two US states and two Spanish provinces) and not in comparison to the LN.

³In another influential paper, Gabaix (1999) has shown that Zipf's law follows as the limiting distribution of an augmented version of Gibrat's law that includes a lower bound for city sizes; also see Gabaix and Ioannides (2004).

⁴In the pure form of Gibrat's law there is no creation of new cities.

3. The overall city size distribution: LN versus DPLN

3.1. Data

The basic data problem for our study is that un-truncated settlement size data, which are needed to fit an entire distribution, are not yet easily available for many countries. What is available are truncated samples of large cities with population size above some threshold level. Such data sets are, for example, used in the cross-country investigation of Zipf's law by Soo (2005) and exist for virtually all countries in the world. It is nevertheless possible to obtain un-truncated settlement size data at least for some countries, and in this paper we present evidence for eight cases. For brevity we mainly focus on Germany, the US, and France, but we additionally include Brazil, Czech Republic, Hungary, Italy and Switzerland in the analysis.

The data for the US is the same that has been used by Eeckhout (2004, 2009) and Levy (2009). It is provided by the US census and includes population sizes for 25,359 settlements ("places") in the year 2000, ranging from 1 to roughly 8m inhabitants in New York City. This data set has two main limitations. First, a "place" is not defined according to economic criteria but follows an administrative definition that, moreover, varies considerably across US states.⁵ An alternative geographical unit are metropolitan statistical areas (MSAs), which are defined in a more meaningful way but are subject to a minimum population size.⁶ Second, the census places, although not subject to a minimum size, do not comprehensively represent the entire US population but only about 74% of it. For Germany the data is provided by the federal statistical office (*Statistisches Bundesamt*). The so-called "DESTATIS" database includes population sizes for 2,075 cities in the year 2006. This data set has comparable problems. A German city is also defined according to administrative boundaries. In addition, the historical awarding of "city rights" is decisive as to whether a settlement is counted as a city or not. The smallest city (Arnis) has 309 inhabitants. Overall, the German data set covers about 72% of the total population in the year 2006. The French data set as provided by the national statistical office (*INSEE*) includes the sizes of 36,674 French administratively defined settlements (communes) in the year 2006. It provides the best coverage as it basically represents the entire French population.

As for the other countries, we consulted the web pages of various national statistical offices (see <http://www.bls.gov/bls/other.htm>) to check for un-truncated settlement size data. In most cases such data are not freely provided, but for the five additional countries mentioned above they are publicly available. Further details about these data can be found in table 1, where we report the number of settlements, the percentage of the overall population that is covered, as well as the minimum and the maximum city size for each country.

3.2. Maximum likelihood estimation of the LN and the DPLN distributions

The first step in the analysis is to fit both the LN and the DPLN parameterizations to the data by using the maximum likelihood (ML) method. Reed and Jorgenson (2005: eq. 28) explicitly derive the log-likelihood function of the DPLN distribution; for the LN this is a standard exercise.

⁵The precise definition of places is explained in the *Geographic Areas Reference Manual* available online under <http://www.census.gov/geo/www/garm.html>

⁶Also see Cuberes (2009) on the pros and cons of administrative versus economic definitions of cities.

Table 1 summarizes the estimated parameters and the corresponding log-likelihoods of the two parameterizations for the eight countries.

TABLE 1 HERE

Notice that the estimated upper tail parameters of the DPLN distribution ($\hat{\alpha}$) are in some cases (France, Czech Republic) very close to unity. This corresponds to an exact validity of Zipf's law. One has to be careful, however, comparing these estimates with Zipf coefficients from the literature, i.e., with shape parameters of a fitted Pareto distribution. This is because the "Zipf coefficients" are highly sensitive to the chosen threshold city size. For example, when running a standard rank-size regression of the type $\log(\text{Rank}) = \log(C) - \zeta \cdot \log(\text{Size})$ for Germany, we estimate $\hat{\zeta} = 1.27$ when including only cities with more than 100,000 inhabitants in the regression, $\hat{\zeta} = 1.34$ with a threshold of 200,000, $\hat{\zeta} = 1.23$ with a threshold of 50,000, and so forth. In other words, a Zipf coefficient exactly equal to unity arises, if at all, only under special assumptions on the minimum city size when fitting a Pareto distribution. It should therefore come as no surprise that the estimates for $\hat{\alpha}$ also deviate from unity. As for the lower tail parameter, there is no focal point to compare to. As can be seen, $\hat{\beta}$ is consistently far greater than unity, but there is no theory saying that the power law in the lower tail should be such that the second-smallest settlement within a country is twice as large as the smallest, or the like. Even when running a "naive" rank-size regression for the lower tail in Germany, $\log(\text{Rank}) = \log(C) + \xi \cdot \log(\text{Size})$, we obtain values that are not even close to unity ($\hat{\xi}=2.48$ for cities smaller than 2,000, $\hat{\xi}=2.09$ for cities smaller than 6,000, and so forth). Generally speaking, an advantage of using un-truncated data is that one does not have to make such arbitrary choices about size thresholds.

We now turn to several informal (visual) and formal tests of the performance of the fitted DPLN versus the simpler but more rigid LN distribution.

3.3. Informal tests

The solid lines in figures 1a-1c show non-parametric kernel density estimations (KDE) of the actual city size distributions in Germany, the US, and France in logarithmic scale using Silverman's optimal bandwidth. The dot-dashed lines in these figures represent the fitted LN, and the dashed lines the fitted DPLN distributions with parameters given above. Upon inspection both parameterizations decently fit the actual distribution in all countries.

In order to address their relative performance, we plot the pointwise vertical differences between the empirical and the two competing theoretical cumulative density functions (cdfs) in the left panels of figures 2a-2c. The right panels show the cumulated deviations at different city sizes. As can be seen, the pointwise differences are larger for the LN than for the DPLN in almost all ranges. A standard Kolmogorov-Smirnov (KS-) test looks at the supremum of these pointwise differences across the entire distribution. This supremum is clearly larger for the LN than for the DPLN in all three countries. Hence, the KS-test would reject the former parameterization earlier than the latter. The cumulated deviations of the DPLN consistently remain below those of the LN, especially in France where we actually have the best relative performance of the DPLN.

FIGURES 2a-2c AND 3a-3c HERE

In figures 3a-3c we plot the empirical cdfs for the three countries with respective confidence bands, which are constructed by using approximations for the critical levels of the 95% KS-test statistics (see Bickel and Doksum, 2001).⁷ The panels on the left refer to the overall cdf and the panels on the right zoom onto the upper tail. For the German case, both theoretical distributions consistently fall inside the 95% confidence band. In other words, *statistically* both distributions cannot be rejected at the 5%-level. For the case of the US, both distributions are sometimes located outside the band in the bottom and medium range, which can be detected in the left panel of figure 2b. However, it can be shown that the LN tends to fall outside that band more often and more clearly than the DPLN. Focussing only on the upper tail (right panel of figure 2b), both the DPLN and the LN are located inside the 95% confidence band throughout. Hence, both parameterizations cannot be rejected in that range of city sizes roughly exceeding $\exp(10) \approx 22,000$ inhabitants. For the case of France the graphical analysis reveals particularly clearly that the DPLN delivers a better fit than the LN, as the latter distribution is actually rejected for a quite wide range of city sizes.

Analogous figures can be provided for the five additional countries for which we have sufficient data, but they are omitted for brevity. They reveal a qualitatively similar picture: The LN fits the data well, and most of the time it cannot be rejected statistically. However, the DPLN delivers a better performance both in the body and in the tails of the distribution.

3.4. Formal tests

Turning now to more formal tests, table 2 condenses the information from figures 2a-2c by integrating up the pointwise vertical differences of the respective theoretical from the empirical distribution. In Germany, for example, the deviations sum up to 15.59 for the DPLN and to 26.94 for the LN (also see right panel of figure 2a), which implies that the LN has 72% higher cumulated deviations. Results look similar for the other countries, i.e., the LN leads to a larger sum of deviations everywhere. The performance difference is particularly strong in France and in the Czech Republic, and smallest in Switzerland and in the US.

From a statistical point of view, the DPLN has a natural advantage as it is the more flexible functional form. We therefore use the log-likelihoods reported in table 1 to compute Akaike's information criterion (AIC) and the related Schwarz criterion (also called "Bayesian information criterion", BIC). Both are model selection criteria that trade-off the precision of a hypothesized distribution and the number of parameters that need to be estimated. Table 2 reports the results. By construction, the distribution with the lower numerical value of the AIC (BIC) is favored. Looking first at the AIC, we find that the values for the DPLN are consistently lower than for the LN distribution in all countries. Turning to the BIC, we obtain a consistent result for seven cases, but for Switzerland the BIC is now in favor of the LN distribution. In the Swiss case the DPLN only leads to a marginally better fit than the LN (the log-likelihoods are almost the same). Since the BIC penalizes the use of additional parameters stronger than the AIC does, the former criterion thus indicates that the simpler model (LN) is sufficient while the latter criterion is still in favor of the richer model (DPLN). For the other seven countries both statistical selection criteria agree that the DPLN is the better suited parameterization.

⁷A similar technique has been used by Eeckhout (2009), who constructs a confidence band for the theoretical (LN) distribution and analyzes if the actual distribution falls inside that band. Our approach of constructing a confidence band for the empirical cdf is useful, because we jointly consider the performance of two theoretical distributions.

TABLE 2 HERE

Given the nested structure of LN and DPLN, we can also compare model performance by a standard likelihood-ratio test. The log-likelihoods are, respectively, denoted by $\ln(L_{LN}^i)$ and $\ln(L_{DPLN}^i)$ for country i , and the test statistic $LR^i = 2 \cdot (\ln(L_{DPLN}^i) - \ln(L_{LN}^i))$ follows the $\chi^2(2)$ -distribution as the DPLN has two parameters more than the LN. As can be seen in table 2, the null hypothesis that the DPLN leads to no significant improvement can be rejected at a very high confidence level (P-value below 1%). The only exception is Switzerland, where we cannot reject the null at the 5%-level. Finally, another approach to model comparison are Bayes factors. This technique is a flexible Bayesian analogue to the likelihood-ratio test, and does not even require one model to be nested in the other. As shown in Kass and Raftery (1995), Bayes factors can be easily approximated by using the Schwarz criterion (BIC). Specifically, to compare the LN and the DPLN distribution we can calculate the Bayes factor for country i as $B^i \approx \exp(S^i)$, where $S^i = \frac{1}{2} (BIC_{DPLN}^i - BIC_{LN}^i)$. The value of B^i can be interpreted by using Jeffrey's scale, and the results in table 2 indicate that there is *strong* evidence in favor of the DPLN. Consistent with our previous results, we find that Switzerland is an exception as the LN is the strongly preferred model for that country.

Summing up, with the exception of Switzerland all model selection criteria clearly show that the DPLN is the better suited model for the true city size distribution, even after being penalized for having more parameters than the LN.

3.5. Rank-size plots for the upper tail

Last, reminiscent of the debate between Levy (2009) and Eeckhout (2009), we analyze the top range in greater detail by using rank-size plots which are a standard tool in the Zipf literature. This final part focuses on the German case for brevity. The dots in figure 4 refer to the actual city sizes of the 100 largest cities (accounting for roughly 27m people or 33 % of the German population) and their respective rank in the national urban hierarchy. The dot-dashed line represents a random sample of the fitted LN distribution, where we rely on 500 iterations. This line indicates how the rank-size plot would look like if the underlying city size distribution were a LN with parameters given above. Similarly, the dashed line represents the sample of the fitted DPLN. The plot on the left is in logarithmic scale. It reveals that the DPLN fits the data very well in the upper tail, which is consistent with the argument by Levy (2009) that the sizes of the largest cities follow a power law. To address the issue raised by Eeckhout (2009), that the low rank logarithm observations lead to a bias in log-log-plots, we provide the same chart in standard scale on the right. This figure leads to essentially the same insight, however.⁸

FIGURE 4 HERE

⁸A similar plot can also be produced for the lower tail of the distribution. Among the 100 smallest cities we also find a distinctive power law pattern that is precisely in line with the predictions of the DPLN distribution. Furthermore we have also conducted an analysis by using log-density-plots, similar as in Eeckhout (2009). That approach also corroborates our findings of the better performance of the DPLN.

It should be noted that these rank-size plots for the upper tail are just one possible method of addressing the goodness of fit of a theoretical distribution. In this paper we have considered several alternative approaches that typically contemplate the *overall* size distribution and not only the upper tail. We thereby followed the notion by Eeckhout (2009) who argues that the focus on the large cities is problematic as the definition of the truncation point is mostly arbitrary.

4. Conclusion

The various methods that have been used in this paper lead to a consistent picture: Although the lognormal (LN) does a good job in fitting the empirical city size distribution across all settlements of a country, the "double Pareto lognormal" (DPLN) distribution does a better job - even after taking into consideration that there are more parameters to be estimated.

Our findings have two main implications. First, they suggest that urban growth across all cities may be better described by the generalized Gibrat process developed in Reed (2002), rather than by the pure form of Gibrat's law. Even though our evidence is indirect, as we do not compare the growth processes directly but the theoretical steady-state distributions, it is consistent with some recent work which also points out that the pure Gibrat's law does not perform well when taking into account all types of settlements (see Michaels et al., 2009). Second, our findings may reconcile the recent debate about city size distributions between Eeckhout (2004, 2009) and Levy (2009) and thereby also build a bridge to the older Zipf literature. The DPLN parameterization implies that city size distributions have a lognormal shape over a wide range, but feature a distinct power law pattern in the tails. These features, in particular the mixture of lognormal with Pareto behavior among large cities, are nicely consistent with the empirical findings by Levy (2009) which have been recently confirmed by Ioannides and Skouras (2009). The urban growth process formalized in Reed (2002) and the resulting asymptotic DPLN distribution may therefore theoretically rationalize those empirical observations.

An issue that is not covered in this paper are the economic microfoundations of urban growth processes. For the pure form of Gibrat's law there already exist economic theories that clarify the foundations for scale-independent urban growth (most notably Eeckhout 2004). The theory by Reed (2002) is still more statistical in nature. It would be interesting to explore which economic forces can give rise to the mechanism of random city formation that is crucial for the Reed-Yule-process. One could, for example, try to extend the Eeckhout-model to allow for an endogenous number of locations by incorporating city birth and death in the style of Henderson (1974). Some recent papers have started, though in a somewhat different context, to explore such questions (e.g., Rossi-Hansberg and Wright 2007), but certainly more work is needed in this area.

References

- [1] Bickel P. and Doksum K., (2001), "Mathematical statistics", Vol. I., Prentice Hall.
- [2] Cuberes, D., (2009), "Sequential City Growth: Empirical Evidence", Working Paper, University of Alicante
- [3] Eeckhout J., (2004), "Gibrat's law for (all) cities", American Economic Review 94, 1429-1451.
- [4] Eeckhout J., (2009), "Gibrat's law for (all) cities: reply", American Economic Review 99, 1676-1683.
- [5] Gabaix X., (1999), "Zipf's law for cities: an explanation", Quarterly Journal of Economics 114, 739-767.
- [6] Gabaix X., Ioannides Y., (2004), "The evolution of city size distributions", in: Henderson, V. and J. Thisse (eds.), Handbook of regional and urban economics, Vol 4. Amsterdam: North-Holland.
- [7] Gibrat R., (1931), "Les inegalites economiques", Paris: Librairie du Recueil Sirey.
- [8] Henderson, V., (1974), "The sizes and types of cities", American Economic Review 64, 640 - 656.
- [9] Ioannides Y., Skouras S., (2009), "Gibrat's law for (all) cities: a rejoinder", Tufts University, Economics Department Discussion Paper.
- [10] Kass R. and A. Raftery (1995), "Bayes factors", Journal of the American Statistical Association 90, 773-795.
- [11] Levy M., (2009), "Gibrat's law for (all) cities: comment", American Economic Review 99, 1672-1675.
- [12] Nitsch V., (2005), "Zipf zipped", Journal of Urban Economics 57, 86-100.
- [13] Michaels G., Rauch F. and Redding S. (2008), "Urbanization and structural transformation", CEPR Discussion Papers 7016, London.
- [14] Mitzenmacher M. (2003) "A brief history of generative models for power law and lognormal distributions", Internet Mathematics 1, 226251.
- [15] Reed, W., (2002), "On the rank-size distribution for human settlements", Journal of Regional Science 42, 1-17.
- [16] Reed, W. and M. Jorgensen, (2005), "The double Pareto-lognormal distribution - A new parametric model for size distribution", Communications in Statistics 34, 1733-1753.
- [17] Rossi-Hansberg E. and Wright M., (2007), "Urban structure and growth", Review of Economic Studies 74, 597-624.
- [18] Soo K., (2005), "Zipf's law for cities: a cross-country investigation", Regional Science and Urban Economics 35, 239-263.
- [19] Yule, U., (1925), "A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S.", Philosophical Transactions of the Royal Society of London, Ser. B 213, 2187.

Figure 1a: City size distribution, Germany 2006

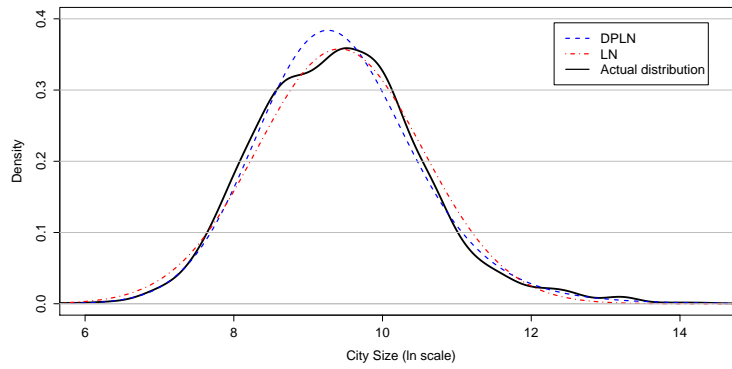


Figure 1b: City size distribution, USA 2000

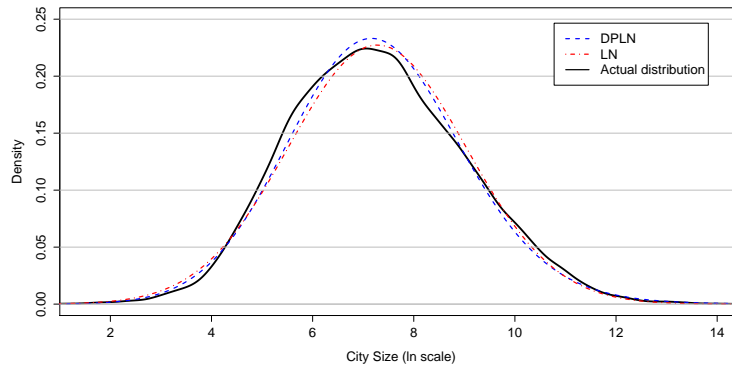


Figure 1c: City size distribution, France 2006

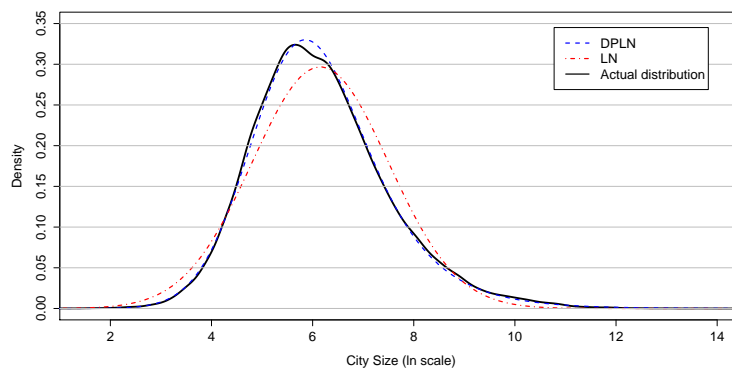


Figure 2a: Deviations of LN and DPLN to the empirical distribution - Germany 2006

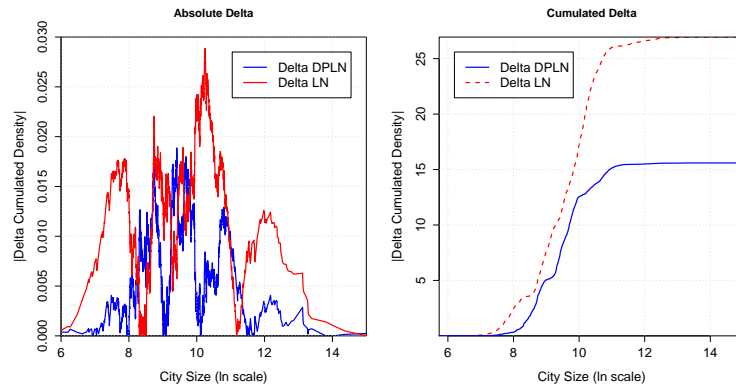


Figure 2b: Deviations of LN and DPLN to the empirical distribution - USA 2000

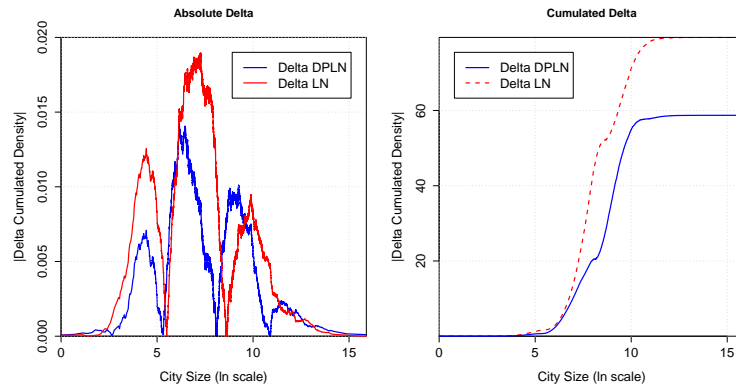


Figure 2c: Deviations of LN and DPLN to the empirical distribution - France 2006

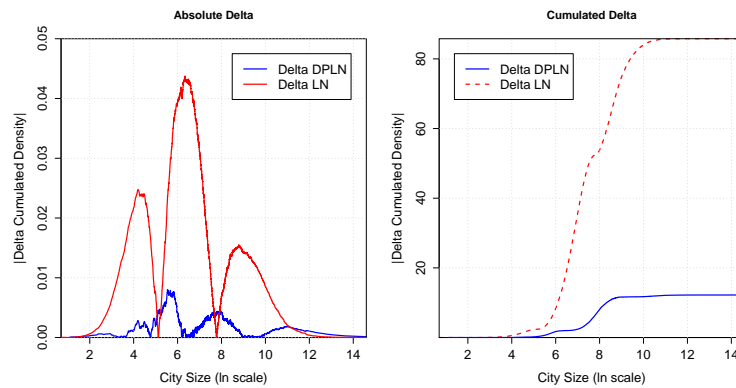


Figure 3a: Empirical city size distribution with 95 %-confidence interval - Germany 2006

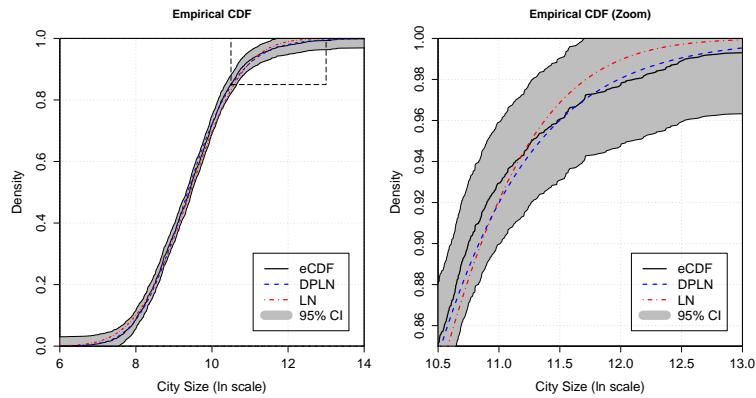


Figure 3b: Empirical city size distribution with 95 %-confidence interval - USA 2000

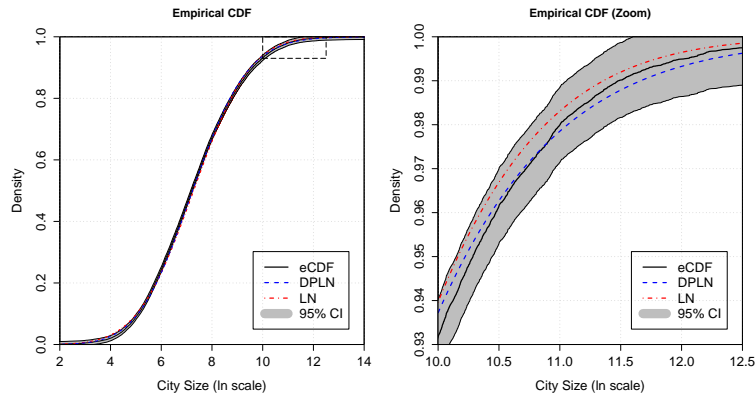


Figure 3c: Empirical city size distribution with 95 %-confidence interval - France 2006

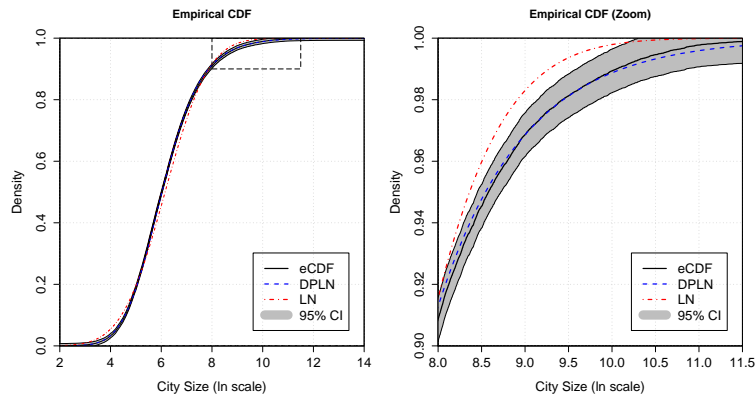


Figure 4: Rank-Size plot for the 100 largest cities - Germany 2006

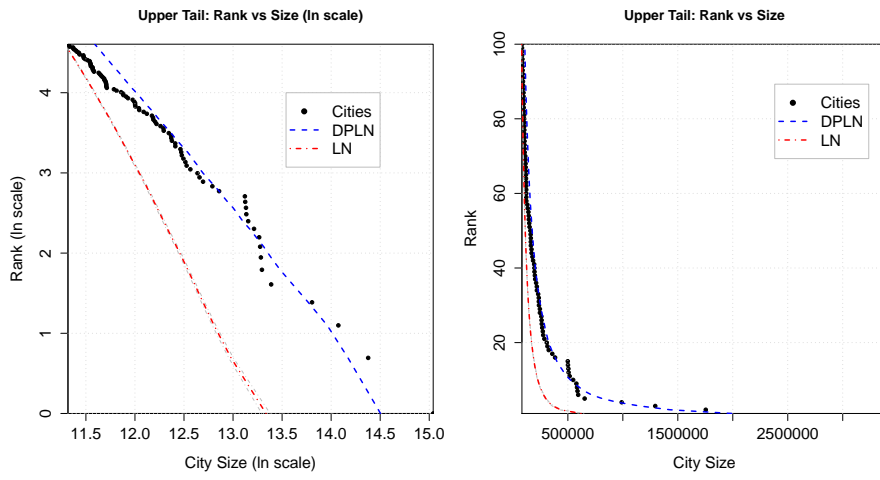


Table 1: Data and estimated parameters of LN and DPLN distribution

	Germany (2006)		United States (2000)		France (2006)		Brazil (2007)	
N	2,075		25,358		36,674		5,564	
Coverage	0.72		0.74		0.97		0.99	
Min	309		1		1		804	
Max	3,404,037		8,008,278		2,181,371		10,886,518	
	LN	DPLN	LN	DPLN	LN	DPLN	LN	DPLN
$\hat{\mu}$	9.425	-	7.277	-	6.152	-	9.384	-
$\hat{\sigma}$	1.114	-	1.753	-	1.344	-	1.137	-
$\hat{\alpha}$	-	1.442	-	1.225	-	1.028	-	1.093
$\hat{\beta}$	-	4.638	-	3.131	-	5.355	-	7.247
$\hat{\mu}_0$	-	8.947	-	6.777	-	5.367	-	8.607
$\hat{\sigma}_0$	-	0.841	-	1.526	-	0.919	-	0.685
$\ln(L_j^i)$	-22,726	-22,687	-234,773	-234,710	-288,484	-287,387	-60,823	-60,536
	Czech (2009)		Hungary (2009)		Italy (2009)		Switzerland (2008)	
N	6,249		3,152		8,101		2,706	
Coverage	0.99		0.99		0.99		0.99	
Min	3		12		33		19	
Max	1,233,211		1,712,210		2,724,347		358,540	
	LN	DPLN	LN	DPLN	LN	DPLN	LN	DPLN
$\hat{\mu}$	6.132	-	6.778	-	7.845	-	6.953	-
$\hat{\sigma}$	1.228	-	1.336	-	1.332	-	1.339	-
$\hat{\alpha}$	-	1.066	-	1.196	-	1.559	-	2.139
$\hat{\beta}$	-	4.818	-	2.070	-	3.691	-	3.287
$\hat{\mu}_0$	-	5.401	-	6.425	-	7.474	-	6.790
$\hat{\sigma}_0$	-	0.766	-	0.922	-	1.135	-	1.215
$\ln(L_j^i)$	-48,467	-48,164	-26,749	-26,702	-77,370	-77,336	-23,442	-23,439

Legend: Un-truncated settlement size data are publicly available from the websites of the respective national statistical offices. See <http://www.bls.gov/bls/other.htm> for a comprehensive list. N is the number of data points (cities) in country i . Coverage is the percentage of the total population in country i and the respective year that is represented by the data set. Min and Max are the population size of the smallest and the largest settlement in the data set. Settlements are classified according to administrative boundaries. See the websites of the national statistical offices for details. Parameters are estimated with the maximum likelihood method. $\ln(L_j^i)$ is the log-likelihood of distribution $j = LN, DPLN$ in the respective country i .

Table 2: Model comparison LN versus DPLN

	Germany (2006)		US (2000)		France (2006)		Brazil (2007)	
	LN	DPLN	LN	DPLN	LN	DPLN	LN	DPLN
Cum. Diff.	26.95	15.59	79.43	58.72	85.78	12.21	140.39	55.38
AIC	45,457	45,382	469,550	469,428	576,971	574,781	121,650	121,079
BIC	45,468	45,404	469,566	469,461	576,988	574,815	121,664	121,106
LR (p-Value)	78.24 (0.01)		126 (0.01)		2194.2 (0.01)		575.2 (0.01)	
Bayes Factor	< 0.0001		< 0.0001		< 0.0001		< 0.0001	
Jeffrey's Scale	Strong for DPLN		Strong for DPLN		Strong for DPLN		Strong for DPLN	
	Czech (2009)		Hungary (2009)		Italy (2009)		Switzerland (2008)	
	LN	DPLN	LN	DPLN	LN	DPLN	LN	DPLN
Cum. Diff.	53.06	6.12	29.19	9.13	38.99	8.34	5.19	3.50
AIC	96,938	96,335	53,501	53,412	154,744	154,680	46,888	46,886
BIC	96,951	96,362	53,514	53,436	154,758	154,708	46,900	46,910
LR (p-Value)	606.5 (0.01)		93.7 (0.01)		67.6 (0.01)		5.92 (0.052)	
Bayes Factor	< 0.0001		< 0.0001		< 0.0001		140.5	
Jeffrey's Scale	Strong for DPLN		Strong for DPLN		Strong for DPLN		Strong for LN	

Legend: The values in the first row report the cumulated vertical deviations between the respective theoretical distribution and the empirical distribution. The Akaike information criterion for country i and distribution j is computed as $AIC_j^i = 2 \cdot k_j - 2 \cdot \ln(L_j^i)$ and the Schwarz criterion as $BIC_j^i = k_j \cdot \ln(N^i) - 2 \cdot \ln(L_j^i)$, with k_j denoting the number of free parameters of distribution j , N^i the number of data points (cities) in country i , and $\ln(L_j^i)$ the log-likelihood as reported in table 1. Both model selection criteria favor the distribution j that yields the lower numerical value. The likelihood-ratio test statistic is calculated according to $LR^i = 2 \cdot (\ln(L_{DPLN}^i) - \ln(L_{LN}^i))$ and follows the $\chi^2(2)$ -distribution. The critical value for a hypothesis test at the 5%-level is equal to 5.99. The Bayes factor for country i is obtained by $B^i \approx \exp(S^i)$, where $S^i = \frac{1}{2} (BIC_{DPLN}^i - BIC_{LN}^i)$. The value of B^i can be interpreted by using Jeffrey's scale (see Kass and Raftery 1995), which implies strong evidence in favor of DPLN if $B^i < 1/10$, moderate evidence if $1/10 < B^i < 1/3$, and weak evidence if $1/3 < B^i < 1$. Values of B^i larger than one indicate evidence in favor of the LN distribution.