Empirical Tests of Zipf's Law Mechanism in Open Source Linux Distribution

T. Maillart,[1] D. Sornette,[1] S. Spaeth,[2] and G. von Krogh[2]

[1]*Chair of Entrepreneurial Risks, Department of Management, Technology and Economics, ETH Zurich, CH-8001 Zurich, Switzerland*
[2]*Chair of Strategic Management and Innovation, Department of Management, Technology and Economics,
ETH Zurich, CH-8001 Zurich, Switzerland*

(Received 30 June 2008; published 19 November 2008)

Zipf's power law is a ubiquitous empirical regularity found in many systems, thought to result from proportional growth. Here, we establish empirically the usually assumed ingredients of stochastic growth models that have been previously conjectured to be at the origin of Zipf's law. We use exceptionally detailed data on the evolution of open source software projects in Linux distributions, which offer a remarkable example of a growing complex self-organizing adaptive system, exhibiting Zipf's law over four full decades.

DOI: 10.1103/PhysRevLett.101.218701 PACS numbers: 89.75.Da, 02.50.Ey, 89.20.Ff

Power law distributions are ubiquitous statistical features of physical, natural and social systems [1,2]. Specifically, the probability density function (PDF) $p(x)$ of some physical variable x, usually a size or frequency, exhibits the power law dependence when

$$p(x) \sim 1/x^{1+\mu} \quad \text{with} \quad \mu > 0. \tag{1}$$

To qualify as a suitable description of a data set, such a PDF should hold within a range $x_{\min} \leq x \leq x_{\max}$ of at least 2–3 decades ($x_{\max}/x_{\min} \geq 10^{2-3}$), and one should understand the origin of the deviations that often appear at both ends $x < x_{\min}$ and $x > x_{\max}$. After claims of universality [3], it is now understood that many different physical mechanisms may be at the origin of power laws in different systems, with possibly widely different exponents μ (see for instance [4–6]).

However, among all power law distributions, one of them, that we refer to as Zipf's law, plays a special role, as it corresponds to the particular value $\mu = 1$, which is at the borderline between converging and diverging unconditional mean $\langle x \rangle$. Historically, Zipf's law described the inverse proportionality between the variable and its rank in a rank-frequency plot [7], which is just another way to state that the distribution of the data follows a power law with the special value $\mu = 1$. Zipf's law has been documented empirically to describe the distribution of the frequency of words in natural languages [7], the distribution of city sizes [8] as well as firm sizes [9–11] all over the world, several distributions characterizing Web access statistics and Internet traffic characteristics [12,13] as well as in bibliometrics, infometrics, scientometrics, and library science (see [14] and references therein). One key challenge is to find and validate the mechanism(s) underlying this universality class $\mu = 1$.

Yule's theory of the power law distribution of the number of species in a genus, family or other taxonomic group [15] and Champernowne's theory of stochastic recurrence equations [16] showed that there are important links between Zipf's law and stochastic growth. On this basis, Simon [17] articulated a simple mechanism for Zipf's law based on Gibrat's law of proportionate effect [18] implemented in a stochastic growth model with new entrants. A modern formulation of Gibrat's law is that growth is a random process, with successive stochastic realizations of the growth rates that are independent of the size of the entity (genera, city, firm, website popularity and so on). This model has recently been rediscovered under the name "preferential attachment" to explain the scale-free networks found in social communities, the World Wide Web, or networks of proteins reacting with each other in biological cells [13,19]. The existence of new entrants in the growth process is one of the additional ingredients complementing Gibrat's law that yields Zipf's law [8,16,20,21]. Gabaix has argued that the special value $\mu = 1$ emerges as a result of the condition of stationarity [8]. Malevergne *et al.* [22] showed recently that Gibrat's law of proportionate growth does not need to be strictly satisfied in the presence of the birth and death of entities following a stochastic growth process: as long as the standard deviation of the growth rate increases asymptotically proportionally to the size and that the average growth rate increases not faster than the standard deviation, the distribution of sizes follows Zipf's law.

However, early on, Mandelbrot confronted Simon in a heated debate over whether the idea of proportional growth has any validity [23]. Surprisingly, the issue is still not settled [4], as proportional growth has not been verified directly in the same systems exhibiting Zipf's law. Here, we empirically verify the constitutive elements entering in the mechanism operating to create the observed universal Zipf's law distribution. For this, we provide an analysis of the growth of packages in open source softwares, as a proxy for the evolution of complex adaptive systems [24]. We study the operating system (Debian Linux). Large Linux distributions typically contain tens of thousands of connected packages, including the operating system and applications, which form a complex web of interdependencies. A measure of the "centrality" of a given package is the number of other packages that call it in their routine, a measure we refer to as the number of in-directed links or connections that other packages have to a given package. We find that the distribution of in-directed links of packages in successive Debian Linux distributions precisely obeys Zipf's law over four orders of magnitudes. We then verify explicitly that the growth observed between successive releases of the number of in-directed links of packages obeys Gibrat's law with a good approximation. As an additional critical test of the stochastic growth process, we confirm empirically that the average growth increment of the number of in-directed links of packages over a time interval Δt is proportional to Δt, while its standard deviation is proportional to $\sqrt{\Delta t}$, as predicted from Gibrat's law implemented in a standard stochastic growth model. In addition, we verify that the distribution of the number of in-directed links of new packages appearing in evolving version of Debian Linux distributions has a tail thinner than Zipf's law, confirming that Zipf's law in this system is controlled by the growth process.

The Linux Kernel was created in 1991 by Linus Torvalds as a clone of the proprietary Unix operating system [25,26], and was licensed under GNU General Public License. Its code and open source license had immediately a strong appeal to the community of open source developers who started to run other open source programs on this new operating system. In 1993, Debian Linux [27] became the first noncommercial successful general distribution of an open source operating system. While continuously evolving, it remains up to the present the "mother" of a dominant Linux branch, competing with a growing number of derived distributions (Ubuntu, Dreamlinux, Damn Small Linux, Knoppix, Kanotix, and so on).

From a few tens to hundreds of packages (474 in 1996 (v1.1)), Debian has expanded to include more than about 18'000 packages in 2007, with many intricate dependencies between them, that can be represented by complex functional networks. Its evolution is recorded by a chronological series of stable and unstable releases: new packages enter, some disappear, others gain or lose connectivity. Here, we study the following sequence of Debian releases: Woody: 19.07.2002; Sarge: 0.6.06.2005; Etch: 15.08.2007; Lenny (unstable version): 15.12.2007; several other Lenny versions from 18.03.2008 to 05.05.2008 in intervals of 7 days.

Figure 1 shows the number of packages in the first four successive versions of Debian Linux with more than C in-directed links, which is nothing but the un-normalized complementary cumulative (or survival) distribution of package numbers of in-directed links. Zipf's law is confirmed over four full decades, for each of the four releases ($x_{\min} = 1$ and $x_{\max} \simeq 10^4$ are the minimum and maximum numbers of in-directed links). Notwithstanding the large modifications between releases and the multiplication of the number of packages by a factor of 3 between Woody and Lenny, the distributions shown in Fig. 1 are all consistent with Zipf's law. It is remarkable that no noticeable cutoff or change of regimes occurs neither at the left nor at the right end-parts of the distributions shown in Fig. 1. Our results extend those conjectured in Ref. [28] for Red Hat Linux. By using Debian Linux, which is better suited for the sampling of projects than the often used SourceForge collaboration platform, we avoid biases and gather unique information only available in an integrated environment [29].

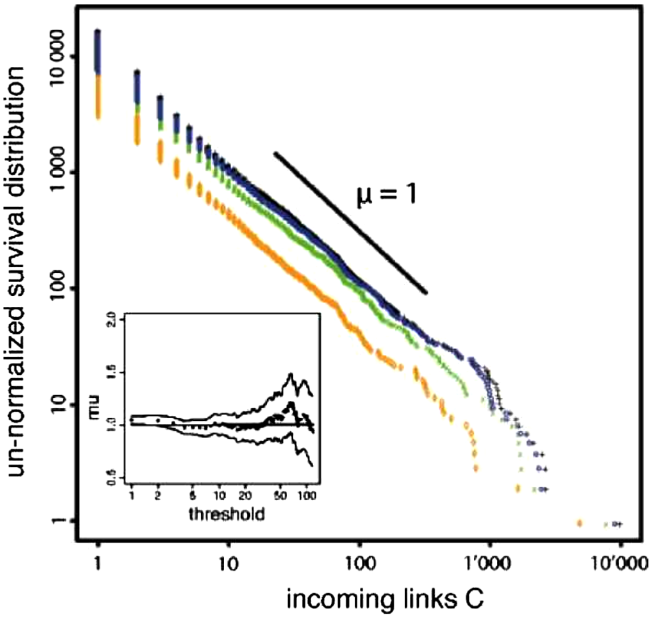

FIG. 1 (color online). (Color Online) Log-log plot of the number of packages in four Debian Linux Distributions with more than C in-directed links. The four Debian Linux Distributions are Woody (19.07.2002) (orange diamonds), Sarge (06.06.2005) (green crosses), Etch (15.08.2007) (blue circles), Lenny (15.12.2007) (black+'s). The inset shows the maximum likelihood estimate (MLE) of the exponent μ together with two boundaries defining its 95% confidence interval (approximately given by $1 \pm 2/\sqrt{n}$, where n is the number of data points using in the MLE), as a function of the lower threshold. The MLE has been modified from the standard Hill estimator to take into account the discreteness of C.

To understand the origin of this Zipf's law, we use the general framework of stochastic growth models, and we track the time evolution of a given package via its number C of in-directed links connecting it to other packages within Debian Linux. The increment dC of the number of in-directed links to a given package over a small time interval dt is assumed to be the sum of two contributions, defining a generalized diffusion process:

$$dC = r(C)dt + \sigma(C)dW, \tag{2}$$

with $r(C)$ is the average deterministic growth of the in-directed link number, $\sigma(C)$ is the standard deviation of the stochastic component of the growth process and dW is the increment of the Wiener process (with $\langle dW \rangle = 0$ and $\langle dW^2 \rangle = dt$ where the brackets denote performing the statistical average). Zipf's law has been shown to arise under a variety of conditions associated with Gibrat's law. The simplest implementation of Gibrat's law writes that both $r(C)$ and $\sigma(C)$ are proportional to C,

$$r(C) = r \times C, \qquad \sigma(C) = \sigma \times C, \tag{3}$$

with proportionality coefficients r and σ obeying the following inequality $r < \sigma$. This later inequality expresses that the proportional growth is dominated by its stochastic component [22]. Accordingly, the heavy tail structure of Zipf's law can be thought of as the result of large stochastic multiplicative excursions. The rest of the Letter is devoted to testing and validating this model.

First, we measure the time evolution of the in-directed links of all packages in the successive Debian releases, by retrieving the network of dependencies following the methodology explained in Ref. [29]. For packages which are common to successive releases, we find that their connectivity, measured for instance by their number C of in-directed links, increases on average albeit with considerable fluctuations. Consider for instance the update from Etch (15.08.2007) to the latest Lenny version (05.05.2008). For each package i which is common to these two versions, we measure the increment ΔC_i of the number C_i of in-directed links to that package from Etch to the latest Lenny version. The left panel of Fig. 2 plots these increments ΔC_i as a function of C_i. This figure is typical of the results obtained on the increments ΔC_i between other pairs of Debian releases. The scatter plot confirms the existence of an approximate proportionality between ΔC_i and C_i, especially for the largest C_i values, in agreement with the first equation of (3). The right panel of Fig. 2 shows the standard deviation of ΔC as a function of C, confirming the second equation of (3). These two panels are nothing but direct evidence of Gibrat's law for package connectivities, which constitutes an essential ingredient of stochastic growth models of Zipf's law [8,16,20,21]. Notice that the large scatter decorating the approximate proportionality between ΔC_i and C_i observed in Fig. 2 and quantified in the right panel of Fig. 2 is an essential ingredient for Zipf's law to appear [22].

We then combine (2) and (3) to predict that, over a not too large time interval Δt, (i) the average growth rate $R(\Delta t) \equiv \langle \Delta C/C \rangle$ should be given by

$$R(\Delta t) = r \times \Delta t, \tag{4}$$

and (ii) the standard deviation of the growth rate

$$\Sigma(\Delta t) \equiv \langle [\Delta C/C]^2 \rangle^{1/2} \tag{5}$$

should be equal to

$$\Sigma(\Delta t) = \sigma \times \sqrt{\Delta t}. \tag{6}$$

This last result derives from the properties of the Wiener process increments dW. We test these two predictions (4) and (6) as follows. Out of the four major Debian releases from 19.07.2002 to 15.12.2007 as well as the several Lenny releases from 18.03.2008 to 05.05.2008 in intervals of 7 days, 66 different time intervals can be formed. For each time interval, we calculate the average growth rate defined by $R(\Delta t) \equiv \langle \Delta C/C \rangle$ and its standard deviation defined by (5). Technically, we estimate $R(\Delta t)$ [respectively $\Sigma(\Delta t)$] as the slope (respectively the standard deviation of the residuals) of the linear regression of ΔC as a function of C. This method allows us to construct confidence bounds by bootstrapping (we reshuffle 1000 times the linear regression residuals). The left [right] panel of Fig. 3 shows the 66 values of $R(\Delta t)$ [$\Sigma(\Delta t)$] as a function of their corresponding time interval Δt (respectively, square-root of Δt),

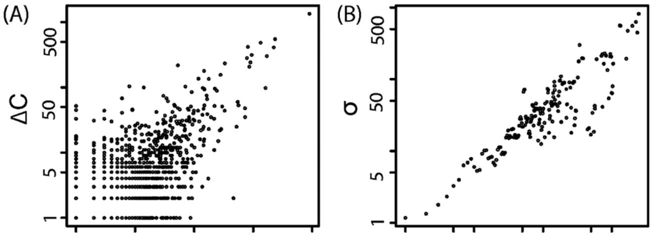

FIG. 2. Left panel: Plots of ΔC versus C from the Etch release (15.08.2007) to the latest Lenny version (05.05.2008) in double logarithmic scale. Only positive values are displayed. The linear regression $\Delta C = R \times C + C_0$ is significant at the 95% confidence level, with a small value $C_0 = 0.3$ at the origin and $R = 0.09$. Right panel: same as left panel for the standard deviation of ΔC.

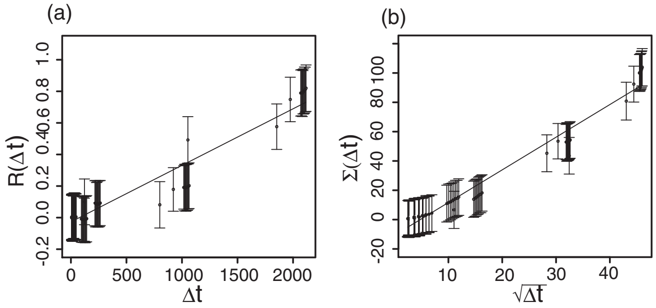

FIG. 3. Dependence of $R(\Delta t)$ and $\Sigma(\Delta t)$ defined, respectively, by $R(\Delta t) \equiv \langle \Delta C/C \rangle$ and (5) as a function of their time interval Δt for the 66 time intervals that can be formed between all the Debian releases in our database (which includes the four major Debian releases from 19.07.2002 to 15.12.2007 as well as the several Lenny releases from 18.03.2008 to 05.05.2008 in intervals of 7 days). The error bars show the 95% confidence intervals, obtained by shuffling 1000 times the linear regression residuals. The straight lines represent the best linear fits. The existence of a genuine linear dependence of R as a function of Δt cannot be rejected ($p < 0.05$) and has a high significance level (square of correlation coefficient $\mathcal{R}^2 = 0.93$). The regression of Σ versus $\sqrt{\Delta t}$ enjoys the same high statistical confidence ($p < 0.05$ and $\mathcal{R}^2 = 0.97$).

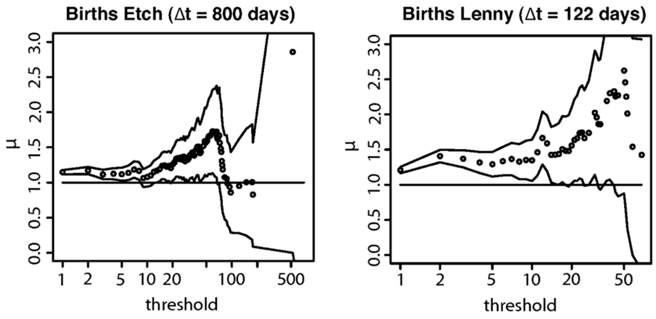

FIG. 4. The right panel shows that the exponent μ of the distribution of C's of new packages appearing between successive unstable Lenny releases separated by one week is a power law with exponent $\mu \simeq 1.5$; the left panel show that the same power law has a smaller exponent closer to 1 as one considers the new packages appearing between two more distant releases. We have verified that this effect is systematic in our database. The exponents μ are obtained by maximum likelihood, adapted to the discreteness of C values. The thin lines defined the 95% confidence intervals.

providing a strong validation of the stochastic growth model (2) and (3).

We now address the question of how the increase of the number of packages interacts with the growth process of the number of links between packages. This issue constitutes an essential ingredient in all the examples where Zipf's law has been documented. Most stochastic growth models based on Gibrat's principle attempt to derive the distribution of sizes directly from the distribution of the size of a single entity as a function of time. Indeed, many models start with the implicit or explicit assumption that the set of entities was born at the same origin of time. This approach is mathematically equivalent to considering that the universe is made only of one single entity. Therefore, the distribution of sizes can reach a steady state if and only if the distribution of the size of a single entity reaches a steady state, which is counterfactual. A more correct model is to take into account the fact that entities do not appear all at the same time but are born according to a more or less regular flow of newly created objects. Competing with the birth process, entities also disappear at a surprisingly high rate. In the context of packages, the evolution of successive Debian releases is indeed punctuated by additions and deletions of many packages. For instance, at the release of the latest stable release (Lenny, 15.12.2007), 885 packages disappeared, partly merged, or were renamed while 2983 packages appeared compared to the precedent release. Clearly, the dynamics of the connectivity between packages depends on the birth as well as demise of packages. Therefore, the stochastic growth model (2) must be supplemented by a model of the birth and death of packages. Such a general model shows that, when volatility dominates over the average growth rate, Zipf's law results from the stochastic growth process and not from the distribution of new entrants' sizes [22].

Figure 4 verifies that the distribution of the numbers C of in-directed links of newly born packages has a tail thinner than Zipf's law, and converges progressively to Zipf's law as the time elapsed between two releases increases, reflecting the increasing impact of the stochastic multiplicative growth process. This confirms that Zipf's law results indeed from the stochastic multiplicative growth process at the level of individual packages in the presence of the birth death of packages.

This project received partial support from the Swiss National Science Foundation Grants 100012-116551 and 105512-106932.

[1] *Fractals in Physics*, edited by A. Aharony and J. Feder, *Proceedings of a Conference in honor of B.B. Mandelbrot, Vence, France* (North Holland, Amsterdam, 1989).

[2] *Proceedings of NATO ASI, Geilo, Norway*, edited by T. Riste and D. Sherrington (Kluwer, Dordrecht, 1991).

[3] P. Bak, *How Nature Works* (Copernicus, NY, 1996).

[4] M. Mitzenmacher, Int. Math. Res. Not. **1**, 226 (2004).

[5] M. E. J. Newman, Contemp. Phys. **46**, 323 (2005).

[6] D. Sornette, *Critical Phenomena in Natural Sciences* (Springer, Heidelberg, 2004), 2nd ed.

[7] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley Press, Cambridge, Mass., 1949).

[8] X. Gabaix, The Quarterly Journal of Economics **114**, 739 (1999).

[9] H. Simon and C. Bonini, Am. Econ. Rev. **48**, 607 (1958).

[10] Y. Ijri and H. A. Simon, *Skew Distributions and the Sizes of Business Firms* (North-Holland, NY, 1977).

[11] R. L. Axtell, Science **293**, 1818 (2001).

[12] L. A. Adamic and B. A. Huberman, Quart. J. Electr. Commerce **1**, 5 (2000).

[13] A.-L. Barabasi and R. Albert, Rev. Mod. Phys. **74**, 47 (2002).

[14] L. A. Adamic and B. A. Huberman, Glottometrics **3**, 143 (2002).

[15] G. Yule, Phil. Trans. R. Soc. B **213**, 21 (1924).

[16] D. Champernowne, Econometrica **63**, 318 (1953).

[17] H. A. Simon, Biometrika **52**, 425 (1955).

[18] R. Gibrat, *Les Inégalités Economiques* (Librarie du Recueil Sirey, Paris, 1931).

[19] A.-L. Barabasi and R. Albert, Science **286**, 509 (1999).

[20] H. Kesten, Acta Math. **131**, 207 (1973).

[21] D. Sornette, Physica (Amsterdam) **250A**, 295 (1998).

[22] Y. Malevergne *et al.*, ssrn.com/abstract=1083962.

[23] B. B. Mandelbrot, Inf. Control **2**, 90 (1959); **4**, 198 (1961); **4**, 300 (1961); H. A. Simon, *ibid.* **3**, 80 (1960); **4**, 217 (1961); **4**, 305 (1961).

[24] C. R. Myers, Phys. Rev. E **68**, 046116 (2003).

[25] Linux Kernel, www.kernel.org.

[26] L. Torvalds, Commun. ACM **42**, 38 (1999).

[27] Debian Linux, www.debian.org.

[28] D. Challet and A. Lombardoni, Phys. Rev. E **70**, 046109 (2004).

[29] S. Spaeth *et al.*, *Proceedings of the 40th Annual Hawaii International Conference on System Science* (IEEE, Piscataway, NJ, 2007), p. 1.