JOURNAL
OF
APPLIED
QUANTITATIVE
METHODS

TOWARD FITS TO SCALING-LIKE DATA,BUT WITH INFLECTION POINTS & GENERALIZED LAVALETTE FUNCTION

Marcel AUSLOOS

E-mail: ausloosm@fastmail.fm

ABSTRACT

Experimental and empirical data are often analyzed on log-log plots in order to find some scaling argument for the observed/examined phenomenon at hands, in particular for rank-size rule research, but also in critical phenomena in thermodynamics, and in fractal geometry. The fit to a straight line on such plots is not always satisfactory. Deviations occur at low, intermediate and high regimes along the *log(x)*-axis. Several improvements of the mere power law fit are discussed, in particular through a Mandelbrot trick at low rank and a Lavalette power law cut-off at high rank. In so doing, the number of free parameters increases. Their meaning is discussed, up to the 5 parameter free super-generalized Lavalette law and the 7-parameter free hyper-generearalized Lavalette law. It is emphasized that the interest of the basic 2-parameter free Lavalette law and the subsequent generalizations resides in its "noid" (or sigmoid, depending on the sign of the exponents) form on a semi-log plot; something incapable to be found in other empirical law, like the Zipf-Pareto-Mandelbrot law. It remained for completeness to invent a simple law showing an inflection point on a <u>log-log plot</u>. Such a law can result from a transformation of the Lavalette law through *x* → *log(x)*, but this meaning is theoretically unclear. However, a simple linear combination of two basic Lavalette law is shown to provide the requested feature. Generalizations taking into account two super-generalized or hyper-generearalized Lavalette laws are suggested, but need to be fully considered at fit time on appropriate data.

Keywords: graphs, plots, nonlinear laws.

1. INTRODUCTION

In recent years, following the rise in the understanding of critical phase transitions [1] through the notion of critical exponents, many results have been presented on log-log graphs. It should be emphasized at once that the search for a straight line fit on such a graph is of interest when the hypothesis of scaling is appropriate for the examined property or effect. Then, the slope on the plot gives some indication of some characteristic exponent at the phase transition because the underlying analytical function, the excess free energy [1], has a homogeneity property. Two other major scientific concepts, related to some underlying scaling hypothesis, have also led to examining log-log plots for various quantities: one is the notion of fractal dimension [2], the other is the rank-size relationship through so called Zipf plots [3].

It is often discussed whether the scaling law should hold over many decades of the *x*-axis variable, -whatever the *x*-axis (reduced temperature ε, bin size *n*, rank *r*, ...). Officially, this "many decades validity" should be the case, if a scaling law fully holds. However, phenomena for which (quasi) straight lines are seen on a log-log plot are rarely found, -outside laboratories or computer simulations. Yet, there is no harm in recognizing that such a straight line existing on a small *x*-axis range indicates the presence of a specific regime; see for example the case of the population size of large italian cities, as illustrated in Fig. 1, for which two regimes rather than a single one can be imagined. Therefore, weak scaling can be accepted as physically suggestive within finite *x*-axis ranges.

Nevertheless, the data can often present convex or concave shapes, and often gaps, jumps, drops (see Fig. 1) or shoulders. Such a large variety of basic shapes demands to pursue some systematic inquiry of the simplest appropriate analytical forms representing complicated data. Much difficulty resides in (interpreting and) theoretically manipulating inflection points, - often visible when a line is drawn through the data "for the eye".

The 2-parameter free power law (using thereafter the discrete variable *r* for the *x*-axis)

$$y_r = \frac{a}{r^{\propto}} \qquad (1)$$

on a log-log plot is referred to Zipf's plot. Zipf had thought that the particular case α= 1 represents a desirable situation, in which forces of concentration balance those of decentralization [3, 4]. Such a case is called the rank-size *rule* [4]-[8]. Thus the scaling exponent *a* can be used to judge whether or not the size distribution is close to some optimum (equilibrium) state.

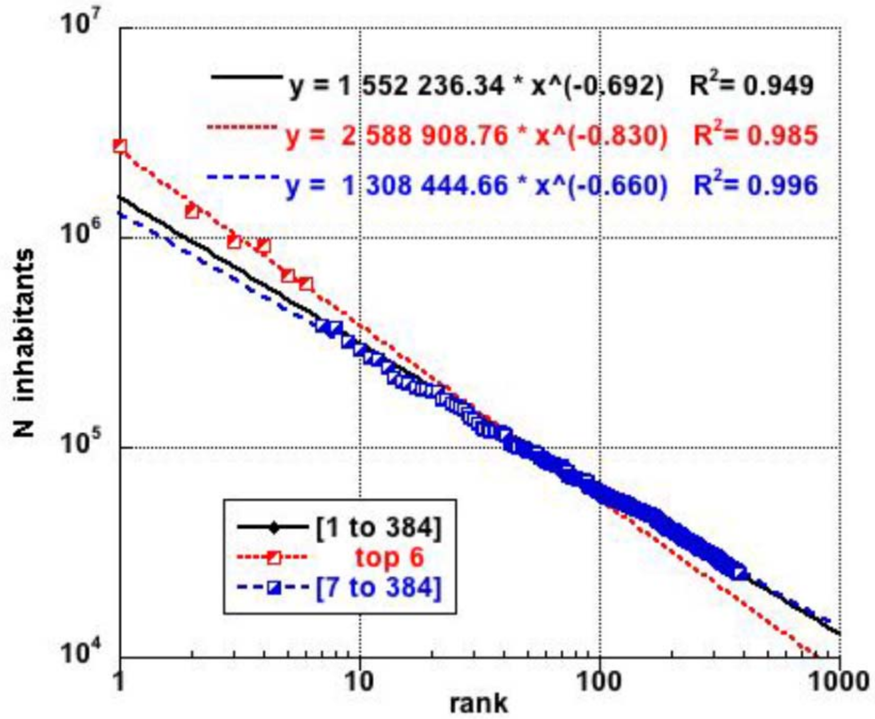

Figure 1: The 384 largest Italian cities ranked by decreasing order of their population size, pointing to a drop after the main 6; different power law fits for the whole range (black line) or when distinguishing two regimes (red and blue line) are indicated with their corresponding correlation coefficient R^2

The pure power-law distribution, for a continuous variable, reads

$$p(k) = \frac{k^{-\gamma}}{\zeta(\gamma)} \qquad (2)$$

where *k* is a positive integer usually measuring some variable of interest; *p(k)* is the probability of observing the value k; γ is the power-law exponent; and $\zeta(\gamma) \equiv \sum_{k=1}^{\infty} k^{-\gamma}$ is the Riemann zeta function; note that γ, in Eq.(2) must be greater than 1 for the Riemann zeta function to be finite.

However, the fit to a straight line on a log-log plot is not always truly perfect, as any reader has surely had the experience considering various data with expected scaling. The error bar (e.g., on γ) can be very large for a R^2 or χ^2 test point of view. Moreover, broadly used methods for fitting to the power-law distribution provide biased estimates for the power-law exponent [9].

The deviations occur in various regimes along the *log(x)*-axis.
When the data crushes at high x-axis value, Lavalette suggested [10] to use the 2-parameter free (κ, χ) form

$$y(r) = \kappa \left[\frac{N\,r}{N-r+1}\right]^{-\chi} \qquad (3)$$

in which the role of *r* as the independent variable, in Eq.(1), is taken by the ratio *r/(N — r* +1) between the descending and the ascending ranking numbers; *N* is the number of data points on the *x*-axis, and $\chi \geq 0$; the +1 role in *(N — r + 1)* is easily understood. Other ways of writing this 2-parameter Lavalette form function are of interest

$$\begin{aligned} \mathrm{y(r)} \;=\; \mathrm{K}\,(N\,\mathrm{r}/(\mathrm{N} — r\,+\,1))^{-\mathrm{X}} &\equiv k\,N^{-\chi}\,(\mathrm{r}/(\mathrm{N} - r\,+\,1))^{-\chi} && (4)\\ &\equiv k\,(N\,r)^{-\chi}\,(\mathrm{N} - r\,+\,1)^{+\chi} && (5)\\ &\equiv \hat{k}\,r^{-\chi}\,(\mathrm{N} - r\,+\,1)^{+\chi}. && (6) \end{aligned}$$

in order to be emphasizing a power law decay with a power law cut-off. The interest in such a function which is strictly decreasing, Fig.2, from infinity at r = 0 under a $r^{-\chi}$ law to a zero value at *r* = *N +1* as $(N - r + 1)^{+\chi}$, best appears on a semi-log plot, Fig. 3: observe the inflection point presence at *r* = *N*/2. The slope s at such a point is equal to $-\,4\,\chi\,\frac{N+1}{N(N+2)}$ which for "large *r*" ~ – 4 χ *(1/N)(1 — 1/N)*. In some sense, it is realistic to reproduce this intermediary regime as y ~ e^{-sr}.

When $\chi \leq 0$, - not a rank-size rule case, the function is increasing, - it is a flipped Lavalette function. Both functions, i.e. with $\chi \geq 0$ or $\chi \leq 0$, are shown in Fig. 2 on a log-log plot, - where the shape is apparently simple, i.e. a power law followed by a sharp cut-off indeed, and on a semi-log plot in Fig. 3, where the shape is "more trivial". On a semi-log plot, Eq.(3) with $\chi \leq 0$, gives a flat N-shape "noid" function (which could be called a "reverse sigmoidal") near its inflection point, which with the correspondingly flat S-shape, but nevertheless called "sigmoid" function, allows to cover various convex and concave data display shapes[1].

[1] Recall that these functions/shapes are found in laboratory when measuring the (I,V) characteristics of junctions or diodes; they present an *N* or *S* shape, beside the Ohm law. The sigmoid or noid form are also describing speculator's different strategies on the stock market [11] .

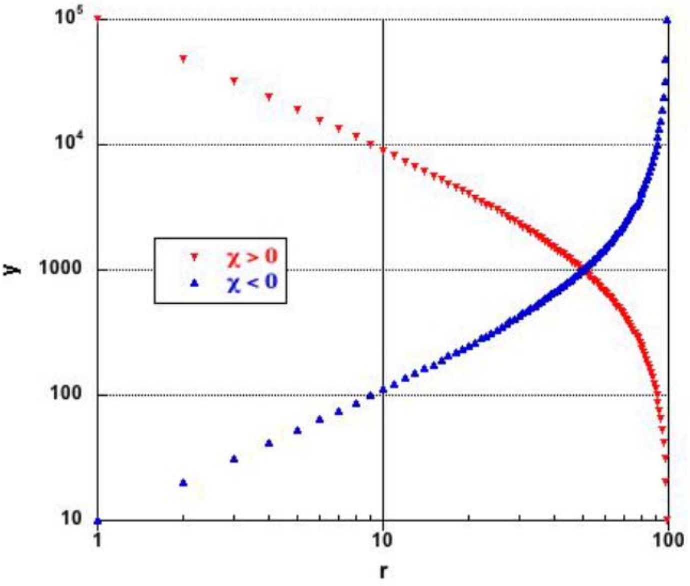

Figure 2: Lavalette function, Eq.(4) with either χ >0 (red dots) or < 0 (blue dots) on a log-log plot, for N = 100 and $\hat{\mathrm{K}}$ = 10^6

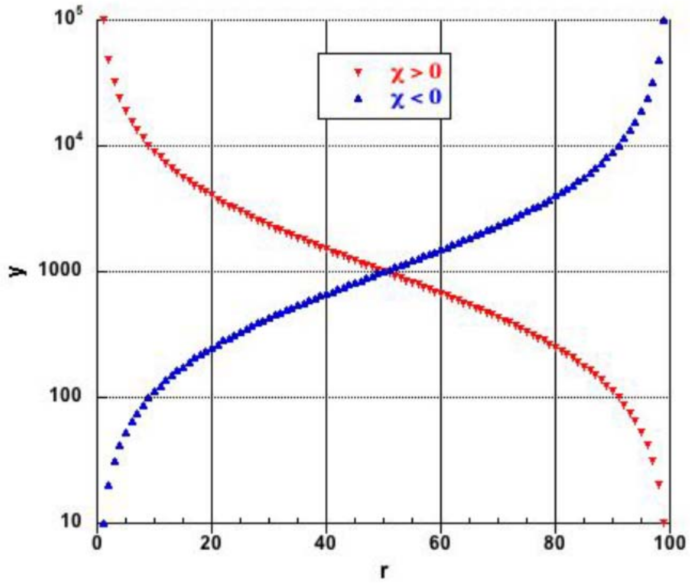

Figure 3: Lavalette function, Eq.(4) with either χ >0 (red dots) or < 0 (blue dots) on a semi-log plot, for N = 100 and $\hat{\mathrm{K}}$ = 10^6, emphasizing the inflection points at $r = N/2$

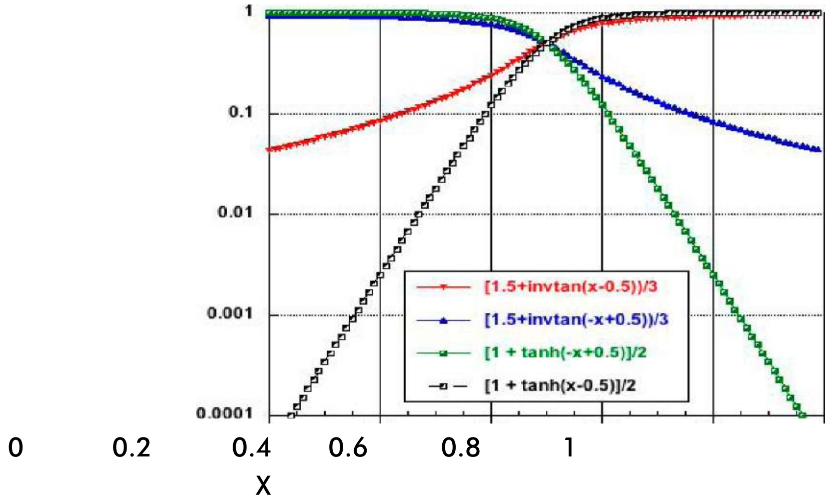

Figure 4: Display of types of sigmoid functions (*invtan*(*x*) and *tanh*(*x*)) on semi-log axes

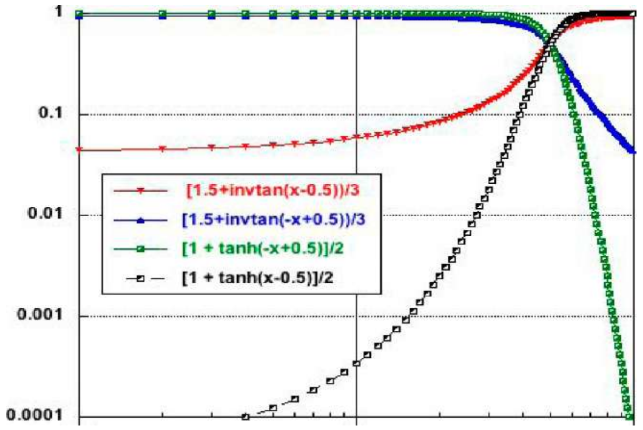

Figure 5: Display of types of sigmoid functions (*invtan*(*x*) and *tanh*(*x*)) on log-log axes

No need to recall that other often seen (or used) 2-parameter free (amplitude and slope at inflection point) have a sigmoid shape; they are *tanh(γx)* and and *invtan(γx)*. There is of course no need to represent such well known functions on classical graphs. They are rarely seen, thus shown on semi-log and log-log plots in Figs. 4-5 respectively. The functions have been adapted and scaled in order to read them on appropriate graphs, for comparison with other functions[2]. A technical point is in order here. Note that *N* (as a factor of *r*, e.g. in Eqs.(3-4) is not really needed. In fact it is more usefully replaced, at fit time, by some simple factor having the order of magnitude of *y(N/2)*. This was made in Fig.6, for example. The Aggregated Income Tax of the 43 cities in the province of Agrigento (AG) in Italy was ranked in decreasing order, for each available year in [2007-2011], from the Italian Minister of Economy, and fitted by an adapted simple Lavalette law, i.e. $\kappa 10^7 \left[r/(43 - r + 1)\right]^{-\chi}$. Note the high regression coefficient values, but a not so visually pleasing fit at high rank.

Finally, considering cut-offs at high rank, there is on the contrary not much discussion in the literature on the wide flattening of the data at high rank, - although such

[2] It should be obvious to the reader that all these *S* or *N* shape functions can occur on different types of plots. The question is whether it can be trivially made *x* → *log*(*x*), whether this "transformation" has any impact on data analysis, and whether some theoretical hypothesis can sustain/justify such a transformation.

cases are encountered, e.g. in co-author ranking [12, 14, 15, 16, 17], and in other "very long flat tail" cases.

For completeness, other 2-parameter free simple functions are recalled in Sect. 5.

2. A FEW 3-PARAMETER FREE FUNCTIONS

Having, introduced well known 2-parameter free functions, to represent complicated data, let us turn on functions with 3 (or more, see below) free parameters, toward elaborating an attempt on how to take into account deviations from simple data approximations by power law-like lines.

2.1. Logistic or Verhulst function

For completeness, recall that the 3-parameter (σ, y_M, and $r_M/_2$) sigmoid forms are well represented through the usually called Verhulst logisitic [18]

$$y(r) = \frac{y_M}{1+e^{-\sigma*(r-\frac{r_M}{2})}} \qquad (7)$$

based on the exponential (growth) function, but invented for limiting the maximum value which such a growth function can reach. This well known function does not need to be shown on an ordinary scale graph. The function is topologically similar to tang(γ x) and invtan(γ). However it is unusual to see this sigmoid function represented or n log-log plot or on semilog plots, whence this is shown in Fig. 7 and Fig. 8, for different σ^* values (with r_M =100), pointing to non-trivial shapes, -also different to those on Fig. 4-5, as the reader can usefully observe by him/ herself.

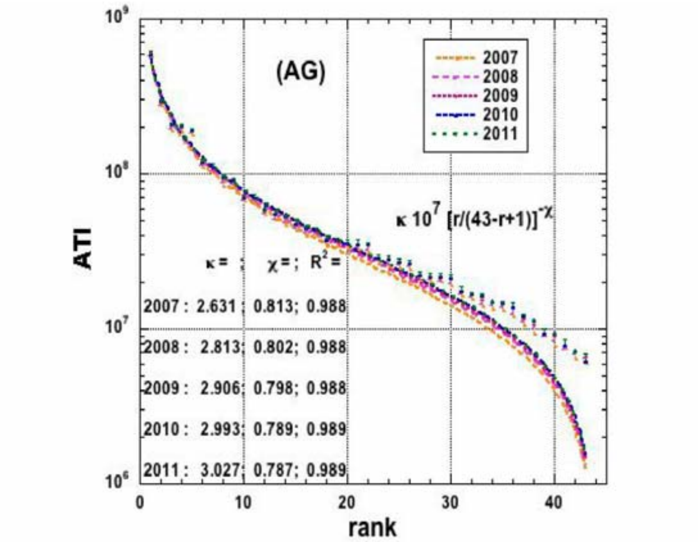

Figure 6: Basic 2-parameter free fit Lavalette law to the Aggregated Income Tax (ATI) of the *N*=43 cities, ranked in decreasing ATI order, in the province of Agrigento, IT, for recent years. Note the high regression coefficient, but not the visually pleasing fit at high rank (*r* ≥ 22)

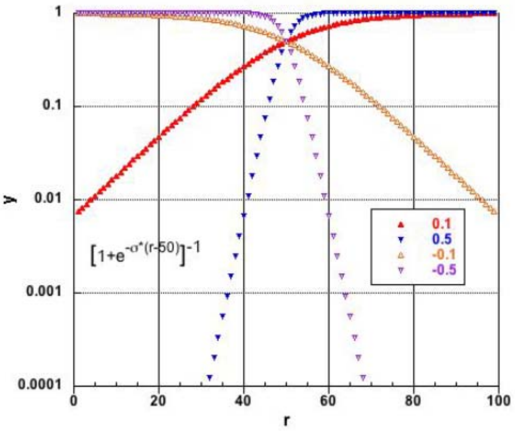

Figure 7: Logistic function, Eq.(7), on semi-log axes

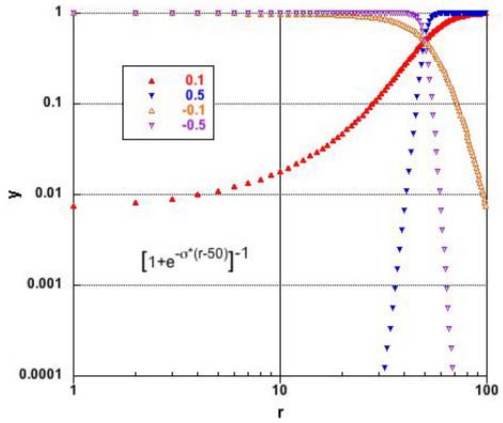

Figure 8: Logistic function, Eq.(7), on log-log axes similar to *tanh(γx)* and *invtan(γx)*. However, it is unusual to see this sigmoid function represented on a log-log plot or on a semi-log plots; whence this is shown, in Fig. 7 and Fig.8, for different σ^* values (with r_M=100), pointing to non trivial shapes, - also different from those on Figs. 4-5, as the reader can usefully observe by himself..

Interestingly, and "obviously", it can noted that some data which could be represented by the Verhulst logisitic, Eq.(7), can be transformed through a simple combination, $[y(r)/(y_M - y(r))]$, into some *Y(r)* which is ≡e^(-σ*(r-r_M/2)). Therefore a semi-log plot of *Y(r)* vs. *r* expectedly leads to a graph with a straight slope from which parameters can be easily deduced [19]; practically, y_M can be used as an appropriate input parameter to optimize the fit.

2.2. Zipf-Mandelbrot function

When the data upsurges at low rank ($r \sim 1$), on a log-log plot, as in [20], one mentions a "king effect" [21], apparently first emphasized in city population size distributions [20]. When the data flattens, below the expected straight line, at low r values, when a so called "queen effect" occurs [12], it is best to modify Eq.(1) into a 3-parameter free form, called the Zipf-Mandelbrot-Pareto (ZMP) law [13], which reads

$$y(r) = \hat{c}/(\eta + r)^{\zeta} \equiv [c/(\eta + r)]^{\zeta}, \qquad (8)$$

since obviously *y(0)* takes a finite value. The value η is understood as a measure of the "harem" [14], - as seen in co-authors of papers distributions.

2.3. Generalized 2-exponent Lavalette function

There is no reason for which the behavior near the crushing point be of (analytically) identical type as the vertical asymptotic behavior at low rank. The basic 2-parameter Lavalette form Eq.(3) can be generalized as a 3-parameter free form [22]

- e.g. allowing two exponents (χ and ξ):

$$y_N(r) = \kappa \frac{(N\,r)^{-\chi}}{(N-r+1)^{-\xi}} \qquad (9)$$

which is emphasizing the number of data points as in Eq.(3), but can be simply written

$$y(r) = \Lambda \frac{[r]^{-\phi}}{[N-r+1]^{-\psi}} \equiv \Lambda [r]^{-\phi} [N+1-r]^{+\psi} \qquad (10)$$

$$\equiv \Lambda [N+1]^{\psi-\phi} \left[\frac{r}{N+1}\right]^{-\phi} \left[1 - \frac{r}{N+1}\right]^{+\psi} \qquad (11)$$

$$\equiv \hat{\Lambda} u^{-\phi} (1-u)^{+\psi} \qquad (12)$$

In fact, the case $\phi > 0$ and $\psi < 0$ is the Feller-Pareto function. The case ϕ = -1 and ψ = +1 is the Verhulst function introduced in the right hand side of the (logistic) evolution differential equation.

However, interestingly, in Eq.(10), both exponents, among the 3-parameters, can take several signs, whence graphical forms can be quite different, as seen in Figs.9-11 shown on the three types of plots.

• but also admitting the same exponent χ, on both tails, but changing the range, leaving free N_1 instead of imposing a predetermined (*N + 1*), - of course imposing N_l - *r* > 0, i.e.

$$y_N(r) = \mathrm{K} \left[\frac{N\,r}{N_1 - r}\right]^{-\chi} \equiv k[N\,r]^{-\chi}[N_1 - r]^{+\chi}, \qquad (13)$$

thus somewhat in the sense of Mandelbrot modification of Zipf law, but at high rank here. In analogy with the theory of critical phenomena [1], one would consider N_l as the "critical range", - analogous to a "critical temperature". One variant of Eq.(13) is merely equivalent to a simple redefinition of κ: $\hat{k} \equiv \kappa N^{-\chi}$. Note again that the role of *N* as a factor of *r* makes "no practical sense". Technically, for optimizing the data fits, it is better to scale the right hand side of such relations, e.g., by a factor 10^m, m obtained, in terms of the order of magnitude of *y*.

3. GENERALIZED 4-PARAMETER FREE LAVALETTE FUNCTION

The modification made in Eq.(13) suggests to apply the Mandelbrot modification also at low rank, in Eq.(9), when there is some flattening of the data at low rank, i.e., one introduces the a similar ZMP trick, as in Eq.(8) on Lavalette function. such that

- combining Eq.(8) idea with the form of Eq.(3), (note that it is different from Eq.(13)), - here keeping the same "names" for the parameters:

$$y_N(r) = k\frac{N^{-\chi}(m+r)^{-\chi}}{(N-r+1)^{-\xi}} \equiv \hat{k}[m+r]^{-\chi}[N-r+1]^{+\xi} \qquad (14)$$

- another 4-parameter free generalized Lavalette function would be

$$y_N(r) = k\frac{N^{-\chi}(r)^{-\chi}}{(N-r+m)^{-\xi}} \equiv \hat{k}[r]^{-\chi}[N-r+m]^{+\xi} \qquad (15)$$

- still a 4-parameter free generalized Lavalette function would be

$$y_N(r) = k\frac{N^{-\chi}(m+r)^{-\chi}}{(N-r+m)^{-\xi}} \equiv \hat{k}[m+r]^{-\chi}[N-r+m]^{+\xi} \qquad (16)$$

Data 23basicFellerPareto

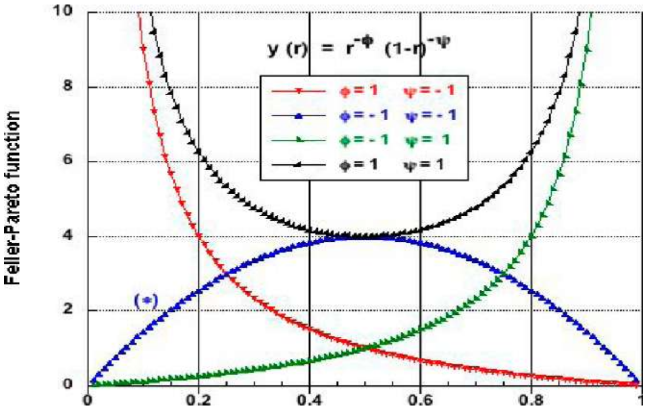

Figure 9: Feller-Pareto function, *y(r)* = $r^{\phi}(1-r)^{-\psi}$, but extended to allow different signs (and possible values) for ϕ and ψ; for readability the amplitude of the ϕ = -1 and ψ = +1 case has been multiplied by a factor 16 as pointed out by (*).

These differ from a generalization [23, 24] based on a Zipf-Mandlebrot function.

4. GENERALIZED 5-PARAMETER FREE LAVALETTE FUNCTION

"Finally", and rather generally a 5-parameter free function is "obviously" in order:

$$y_N(r) = k\frac{N(m+r)^{-\chi}}{(N-r+n)^{-\xi}} \equiv \hat{k}[m+r]^{-\chi}[N+n-r]^{-+\xi} \qquad (17)$$

No graph illustrates this super-generalization; a simple combinatory calculation indicates that one would ask for ten of them. It is better to suggest to envisage such a form when those with a lower number of free parameters do not lead to satisfactory or successful fits. It seems that one can rather easily understand the effect of the new parameters when examining the functions.

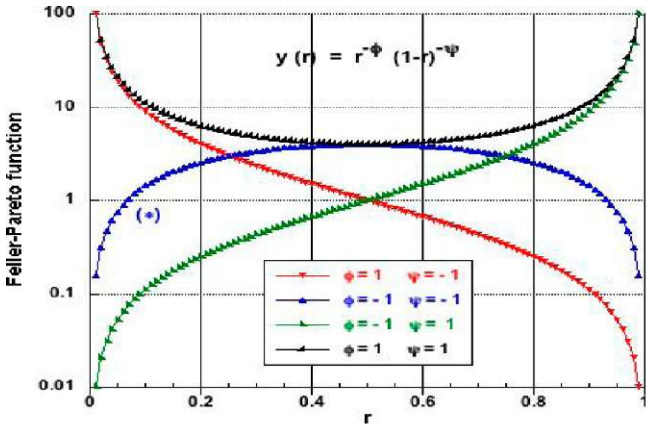

Figure 10: Feller-Pareto function, *y(r)* = $r^{\phi}(1-r)^{-\psi}$, on a semi-log plot, but extended to allow different signs (and possible values) for ϕ and ψ; for readability the amplitude of the ϕ = -1 and ψ = +1 case has been multiplied by a factor 16 as pointed out by (*)

5. A FEW OTHER FORMULAE FOR FITS

For completeness, recall a few other often used formulae for fitting data (often) on log-log plots.

5.1. 2 parameters

Beside the power law, Eq.(1) and the basic 2-parameter Lavalette form Eq.(3), one should mention

- the (2 parameter) exponential case

$$y(r) = b\, e^{-\beta r} \qquad (18)$$

- a law suggested by Tsallis and de Albuquerque[3] (for ranking paper citations) [25]

$$y(r) = \frac{\phi}{[1+(\psi'-1]\ln(r))^{\psi}} \qquad (19)$$

with $\psi' \equiv \psi$, although there does not seem any reason why it should be so.

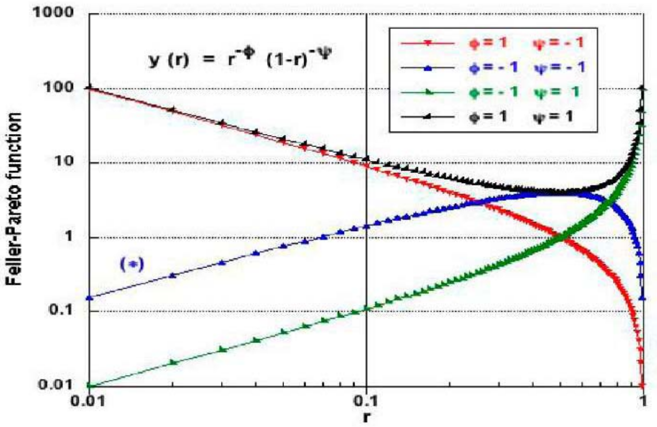

Figure 11: Feller-Pareto function, *y(r)* = $r^{\phi}(1-r)^{-\psi}$ on a log-log plot, but extended to allow different signs (and possible values) for ϕ and ψ; for readability the amplitude of the ϕ = -1 and ψ = +1 case has been multiplied by a factor 16 as pointed out by (*)

[3] correcting a misprint in [23].

- the log-normal distribution [26],

$$y(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right) \qquad (20)$$

where $x > 0$, μ and σ are the parameters, mean and standard deviation of the log of "variable" in the data distribution.

5.2. 3 parameters

Beside the Verhulst logistic form, Eq.(7) and the Zipf-Mandelbrot-Pareto (ZMP) law [13], Eq.(8), other often used 3-parameter statistical distributions, generalizing the power and/or exponential law are to be examined :

- the Yule-Simon distribution, i.e. a power law with exponential cut-off [27] (the free parameters are: d, α, and λ)

$$y(r) = d\ r^{-\alpha} e^{-\lambda r} \qquad (21)$$

- the stretched exponential [21] (the free parameters are: $\theta, \mu\ and\ \nu$)

$$y(r) = \theta\ r^{\mu-1} e^{-\nu\, r^{\mu}} \qquad (22)$$

- the Gompertz double exponential [28] (the free parameters are: g_1, r_2, and g_3)

$$y(r) = g_1 e^{-e^{-(r-r_2)/g_3}} \qquad (23)$$

These functions also bend in convex form on a log-log plot.

5.3. 4 parameters

There are several possible generalizations of the above, often introducing the Mandelbrot trick, at low rank, i.e. $r \rightarrow r + \rho$, with a possibly different ρ at high and low ranks, but they do not seem of major interest. Indeed, look at

- a ZMP4 form, e.g.,

$$y(r) = m_3/(m_2 + m_4\, r)]^{\zeta}, \qquad (24)$$

which obviously reduces to Eq.(8) by a trivial change in the parameter notations, e.g. $\widehat{m_3} \rightarrow m_3/m_4^{\zeta} \equiv c$, and $m_2/m_4 \equiv \eta$,

- or

$$y(r) = m_3(r - m_4)^{-m_1} e^{-m_2(r-m_4)} \qquad (25)$$

with $m_4 \equiv$ to some r_0, which it is nothing else that

$$y(r) = \widehat{m_3}(r - m_4)^{-m_1} e^{-m_2 r} \qquad (26)$$

Usually such functions reproduce one tail but not the other. Technically, such improvements do not change in a dramatic way the regression coefficient, since the high rank tail does not have a great impact upon this coefficient, - because of the change in the order of magnitude between the low and high rank regions.

6. HYPERGENERALIZED (LAVALETTE) FIT FUNCTIONS

It might be reminded that the modification of Keynes differential growth equation by Verhulst through a *(1 - x)* term was purely a mathematical *ad hoc* mean to avoid a full exponential growth. There is no economic or demographic argument to use a linear *(1 - x)* term; a quadratic term *(1 - x^2)* or any other polynomial decaying near *x = 1* or many more complicated terms could be used. Therefore,considering that the basic phenomena might not necessarily depend linearly on *r*, but the rank-size rule should (or could) contain higher order terms, other generalizations may come in mind within the present considerations. One such a case was found in considering city sizes (in Bulgaria, e.g. [29, 30]), but might occur more frequently than "expected", - however are likely not reported because of missing framework. Therefore, a hypergeneralization of Lavalette function can be imagined:

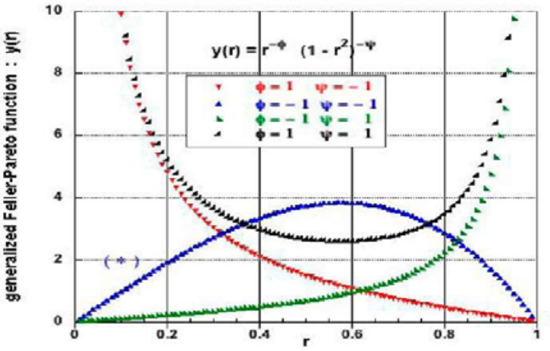

Figure 12: Hypergeneralized Feller-Pareto function, y(r) $= r^{\phi}(1 - r^2)^{-\psi}$, on ordinary axes; (*) indicates that the function has been multiplied by a factor 16 for better readability

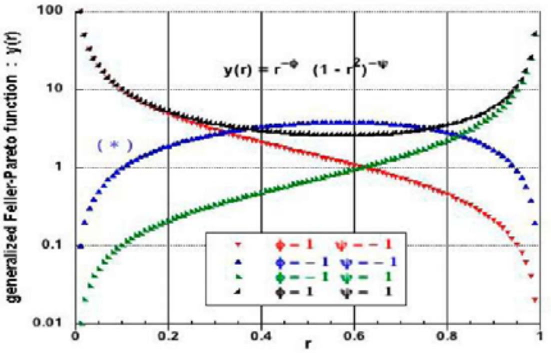

Figure 13: Hypergeneralized Feller-Pareto function, y(r) $= r^{\phi}(1 - r^2)^{-\psi}$, on a semi-log plot; (*) indicates that the function has been multiplied by a factor 16 for better readability

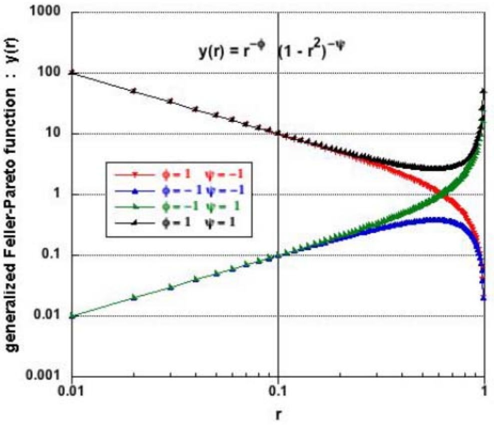

Figure 14: Hypergeneralized Feller-Pareto function, y(r) $= r^{\phi}(1 - r^2)^{-\psi}$, on a log-log plot

- the 3-parameter generalized Lavalette form [22] can be hypergeneralized into

$$y(r) = \frac{[\Lambda\, r^n]^{-\phi}}{[N+1-\ r^m]^{-\psi}} \quad or \quad = \Lambda \frac{[r^n]^{-\phi}}{[N+1-\ r^m]^{-\phi}} \qquad (27)$$

- the 4-parameter generalized Lavalette form [24] can be hypergeneralized into

$$y(r) = \frac{[\Gamma\,(r^n+\nu)]^{-\eta}}{[N-r^m+\nu]^{-\zeta}} \quad or \quad = \Gamma \frac{[r^n+\nu]^{-\eta}}{[N-\ r^m+\nu]^{-\zeta}} \qquad (28)$$

- the 5-parameter supergeneralized Lavalette form (also) can be hypergen - eralized into

$$y(r) = \frac{[\Gamma\,(r^n+\mu)]^{-\eta}}{[N-r^m+\nu]^{-\zeta}} \quad or \quad = \Gamma \frac{[(r^n+\mu)]^{-\eta}}{[N-\ r^m+\nu]^{-\zeta}} \qquad (29)$$

Note that variants : $[(r^n + \nu)] \rightarrow [(r + \nu)^n]$, and $[(r^n + \mu)] \rightarrow [(r+\mu)^n]$, with or without $[(r^m - \nu)] \rightarrow [(r - \nu)^m]$, can be written. The writing choice is left for fit optimization

7. ON INFLECTION POINTS ON LOG-LOG PLOTS

Finally, not the least, the above formulae have much emphasized possible fits which indeed allow inflection points on semi-log graphs, but have left opened the case of inflection points on log-log graphs. Let it be understood that such a case occurs when some power law decay ("from infinity") at low rank is followed by another intermediary regime before some cut-off occurs at high rank. A trivial transformation x → *log(x)* of all the above formulae is possible, but demands much reflection. Indeed, one could transform the basic Lavalette equation to read

$$y(r) \simeq \left[\frac{N\log(r)}{N+1-\log(r)}\right]^{-\chi} \qquad (30)$$

and similarly all others. But it remains to be done some interpretation and much theoretical work !

Another possibility comes from realizing that if there is an inflection point, the slope has the same (negative) sign for the whole *r* range, but the derivative of the slope has some structure, i.e.allowing for a concave to a convex shape of the approximation to the data. The intermediary regime can also be considered in a first approximation to be a scaling law. The high rank regime can be either a Lavalette cut-off or an exponential cut-off. Therefore the following functions can be appropriately imagined

- in its most generalized form, with power law cut-off

$$y(x) = [A(x+m_5)^{-m_1} + B(x+m_6)^{-m_2}](N+m_4-x^{m_7})^{m_3} \qquad (31)$$

- or with an exponential cut-off

$$y(x) = [A(x+m_5)^{-m_1} + B(x+m_6)^{-m_2}]e^{-m_3\,(x+m_4^{m_7}} \qquad (32)$$

A few of such cases are shown in Figs.21-22 demonstrating the interest of such forms in order to discuss inflection points on log-log plots.

8. APPLICATIONS

This section serves as an illustration of a few cases discussed above, displaying some data on either semi-log or log-log plots for comparison. However the data pertains to some empirical study requesting a brief introduction. In so doing, it is hoped that the "universality" of the approach receives a positive argument.

Consider the following investigation. In Italy, 638 cities contain a saint or an angel name, as counted after translating the names into italian, from french, german, or local dialects (like Santu Lussurgiu = Santo Lussorio, or Santhia who is Santa Agata), Note that Sant'Angelo (24 times), San Salvatore (5 times) or Santa Croce (7 cities), and similar "concepts" (Sansepolcro) are not counted. Some distinction can be made between male and female saints. Note that two cities have a name with two saints. The name of the saints can be ranked according to their frequency [31] and an appropriate statistical analysis can follow for the rank-frequency distribution.

However, one can also ask, as did Pareto in 1896, how many times one can find an "event" greater than some size *n*, i.e. study the *size-frequency relationship*. Pareto found out that the cumulative distribution function (CDF) of such events follows an inverse power of *n*, or in other words, *P [N > n]* $\sim n^{-\omega}$., - whence the frequency *f* of such events of size *n*, (also) follows an inverse power of *n*.

Thus, one can count how many cities have a happax hagionym, how many cities have a name with a saint occurring only twice, etc. up to how many cities have a name associated to the "most popular" (= most frequent) saint (San Pietro). This counting is normalized and turned into a probability distribution, i.e. *CDF(n)*. The data is illustrated in Figs. 15-20, either with semi-log or log-log plots, and fits with a Zipf-Mandelbrot or Lavalette function.

Short final comments: (i) two "queen effects" and a "king effect" are well seen on Fig. 16; (ii) the *CDF* shows a pronounced cut-off at high *n* in all cases. Therefore, it could be argued that the *CDF* is less pertinent to observe minute effects. This is understandably true, since the *CDF* results from an integration scheme. However, again understandably, the *CDF* fits are much more stable. No need to say that one should not report too precise parameter values, since these are non linear fits; a final technical information: the Levenberg-Marquardt algorithm was used.

9. CONCLUSIONS

It has been shown that semi-log plots are of interest in order to analyze whether experimental or empirical data are underlined by some scaling argument for the observed/examined phenomenon at hands. The fit to a straight line on log-log plots is not always satisfactory indeed. Deviations occur at low, intermediate and high regimes along the x-axis. Several improvements of the mere power law fit have been discussed, in particular through a Mandelbrot trick at low rank and a Lavalette power law cut-off at high rank.

In so doing, the number of free parameters increases. Their meaning has been discussed, up to the 5 parameter free super-generalized Lavalette law and the 7-parameter free hyper-generealized Lavalette law[4] . It has been emphasized that the interest of the basic 2-parameter free Lavalette law and the subsequent generalizations resides in its "noid" (or sigmoid, depending on the sign of the exponents) form on a semi-log plot; something incapable to be found in other empirical law, like the Zipf-Pareto-Mandelbrot law. The connection with other laws, e.g. Feller-Pareto and Verhulst logistic laws, as been pointed out.

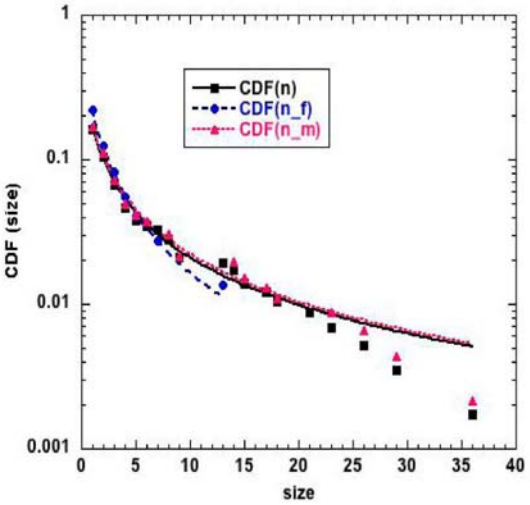

Figure 15: Semi-log plot of the cumulative distribution function (CDF) of the frequency of Italian cities containing a saint name n-times, so called "size", given according to the Zipf-Mandelbrot-Pareto function, like Eq.(8), distinguishing between male (*n_m*) and female (*n_f*) saint names; the fit parameter values are given in Fig.16. Observe the need for a cut-off at high rank/size.

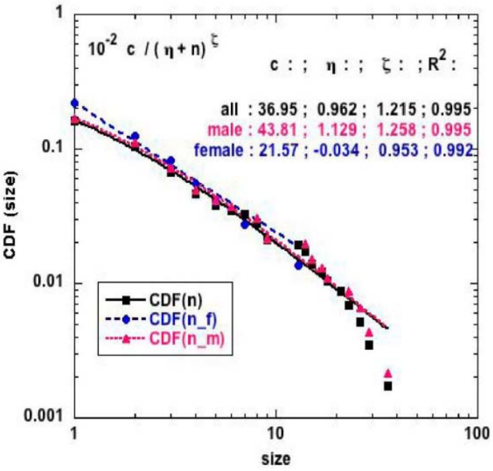

Figure 16: Log-log plot of the cumulative distribution function (CDF) of the frequency of Italian cities containing a saint name n-times given according to the Zipf-Mandelbrot-Pareto function, like Eq.(8), distinguishing between male (*n_m*) and female (*n_f*) saint names; observe that η is negative for the female case, pointing to a king effect (Santa Maria), and queen effects, since $\eta \geq 0$, for the males and the overall distribution. Observe the need for a cut-off at high rank/size.

[4] In this conclusion, one could recall that 6 or 7-parameter free functions are also used for fitting data like in financial market crash predictions [32, 33, 34, 35, 36] or in earthquake predictions [37]

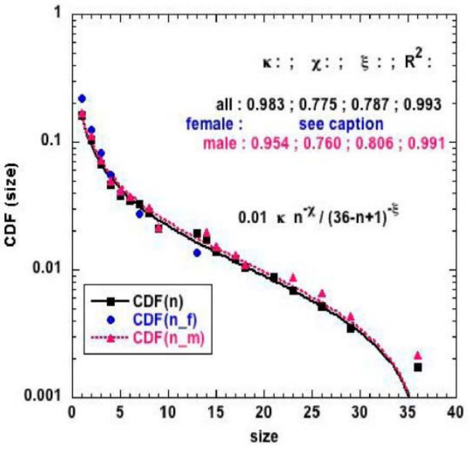

Figure 17: Semi-log plot of the cumulative distribution function (CDF) of the frequency of Italian cities containing a saint name, given *n*-times, so called "size"; fit according to a Lavalette function with 3 free parameters, Eq.(10), for the distribution of all such 36 cities(black line) or only those 36 with a male saint name (n_m; red line); the parameter values for the female case are given in Fig.18, with the corresponding fit. Observe the interest of leaving the high rank/size value be a free parameter, as on Fig.19

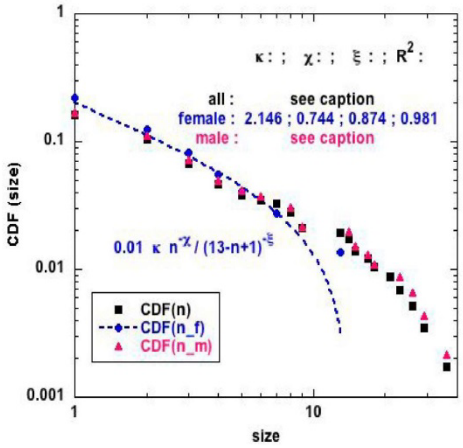

Figure 18: Log-log plot of the cumulative distribution function (CDF) of the frequency of Italian cities containing a saint name given *n*-times, so called "size"; fit according to a Lavalette function with 3 free parameters, Eq.(10) is shown for the distribution of only those 13 cities with a female saint name (n_f; blue line); the parameter values for the male case and the whole distribution are given in Fig.17, with the corresponding fits. Observe the interest of leaving the high rank/size value be a free parameter, as on Fig.20.

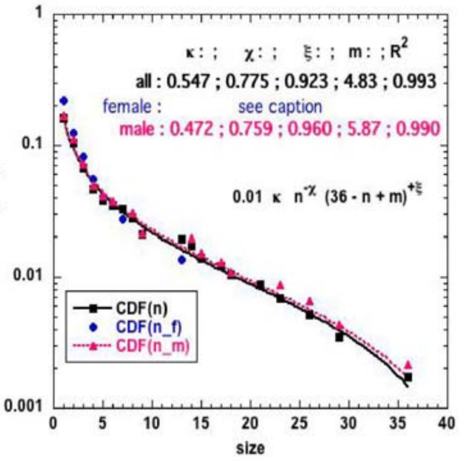

Figure 19: Semi-log plot of the cumulative distribution function (CDF) of the frequency of Italian cities containing a saint name given *n*-times, so called "size"; fits with a 4 parameter free Lavalette function, Eq.(15) are shown for the distribution of all such 36 cities (black line) or only those 36 with a male saint name (n_m; red line); the parameter values for the female case are given in Fig.20, with the corresponding fit.

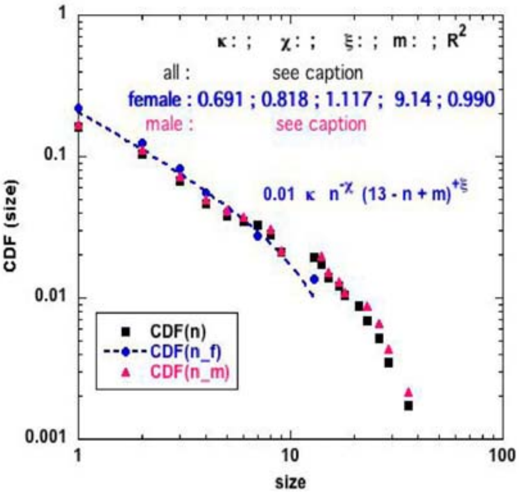

Figure 20: Log-log plot of the cumulative distribution function (CDF) of the frequency of Italian cities containing a saint name given *n*-times, so called "size"; fit according to a Lavalette function with 4 free parameters, Eq.(15) shown for the distribution of only those 13 cities with a female saint name (n_f; blue line); the parameter values for the male case and the whole distribution are given in Fig.19, with the corresponding fits.

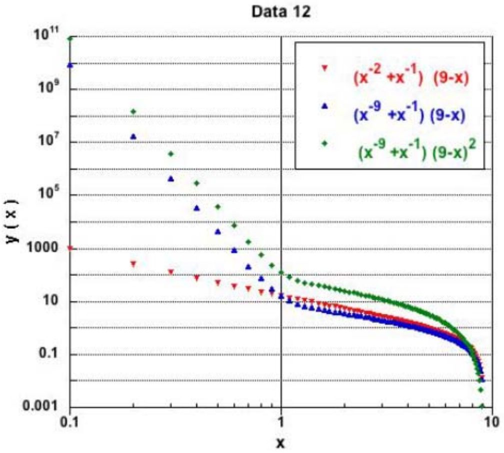

Figure 21: Display of a "simple" function with inflection point on a log-log plot, allowing for fit to data with large king or queen effect and power law cut-off, i.e. with an inflection point in the middle range, as approximated by a simple function for which the general form is Eq.(31).

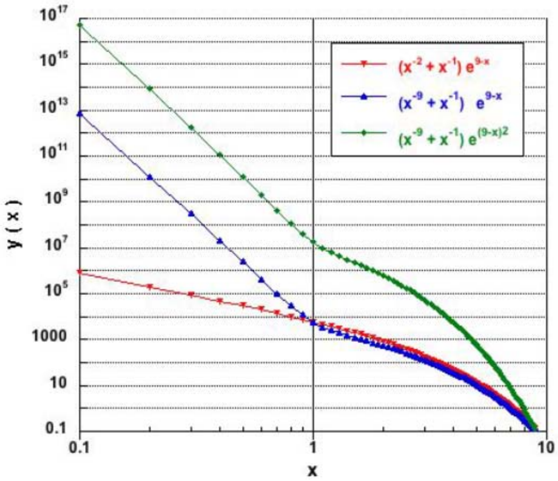

Figure 22: Display of a "simple" function with inflection point on a log-log plot, allowing for fit to data with large king or queen effect and exponential cut-off, i.e. with an inflection point in the middle range, as approximated by a simple function for which the general form is Eq.(32).

It has been shown that the additional parameters introduced into the basic Lavalette function, Eq.(3), facilitates a rather good reproduction of rank-probability distribution in the ranges of small and high rank values. Indeed, each parameter or ratio involved in the suggested modification of Lavalette function, Eq.(3), enhances the fit in different ranges of *r*.

It has remained for completeness to invent a simple law showing an inflection point on a log-log plot. Such a law could have been the result of a transformation of the Lavalette law through *x* → *log(x)*, but this meaning is theoretically unclear. It has been shown that a simple linear combination of two basic Lavalette law provides the requested features.

Generalizations taking into account two super-generalized or hyper-generalized Lavalette laws are suggested, but need to be fully considered at fit time on appropriate data.

A few examples are used for illustrating various points, like deviations or visually unattractive fits, - though the regression coefficient R^2 is often quite satisfactory looking. Examples have been taken mainly for rank-size rule research. However, in order to demonstrate a larger validity of generalizing the usual fit formulae, and some interest for generalizing the basic concepts, some short analysis has been presented of the cumulative distribution function (CDF) of the city names in Italy containing a (male or female) saint name.

REFERENCES

1. H. E. Stanley, **Phase Transitions and Critical Phenomena**, Oxford Univ. Press, Oxford, 1971

2. B.B.Mandelbrot, **The Fractal Geometry of Nature**, W.H.Freeman, New York, 1982

3. Zipf, G.K. **Human Behavior and the Principle of Least Effort : An Introduction to Human Ecology**. Cambridge, Mass.: Addison Wesley, 1949.

4. X. Gabaix, **Zipf's law for cities: An explanation**, The Quarterly Journal of Economics **114,** 739-767, 1999.

5. G. Brakman, H. Garretsen, C. van Marrewijk, and M. van den Berg, **The Return of Zipf: Towards a Further Understanding of the Rank-Size Distribution**, Journal of Regional Science **39,** 182-213, 1999.

6. D.R. Vining, Jr., **The rank-size rule in the absence of growth**, Journal of Urban Economics **4,** 15-29,1977.

7. J.-C. Cordoba, **On the distribution of budget sizes**, Journal of Urban Economics **63,** 177-197, 2008.

8. Yanguang Chen, **The rank-size scaling law and entropy-maximizing principle**, Physica A **391,** 767-778, 2012.

9. Michel L. Goldstein, Steven A. Morris, and Gary G. Yen, **Problems with Fitting to the Power-Law Distribution**, *0402322f ittingtopowerlaw_v3?*

10. D. Lavalette, **Facteur d'impact: impartialite ou impuissance?**, Internal Report, INSERM U350, Institut Curie, France, Nov. 1966

11. M. Ausloos, **Gas-kinetic theory and Boltzmann equation of share price within an equilibrium market hypothesis and ad hoc strategy**, Physica A 284, 385-392, 2000.

12. Ausloos, M. **A scientometrics law about co-authors and their ranking**. The co-author. *Scientometrics,* 95(3), 895-909, 2013.

13. R. A. Fairthorne, **Empirical hyperbolic distributions (Bradford-Zipf- Mandelbrot) for bibliometric description and prediction**, Journal of Documentation 25, 319-343, 1969.

14. Ausloos, M. **Binary Scientific Star Coauthors Core Size**. Scientometrics, in press, 2014

15. Bougrine, H. **Subfield Effects on the Core of Coauthors**. Scientometrics, 98(2), 1047-1064, 2014.

16. Miskiewicz, J. **Effects of Publications in Proceedings on the Measure of the Core Size of Coauthors**. Physica A, 392(20), 5119-5131, 2013.

17. G. Rotundo, **Black-Scholes-Schrdinger-Zipf-Mandelbrot model framework for improving a study of the coauthor core score**, Physica A, in press DOI information: 10.1016/j.physa.2014.02.011

18. Verhulst, P.F., **Recherches mathematiques sur la loi d'accroissement de la population**, Nouveaux Memoires de l'Academie Royale des Sciences et Belles-Lettres de Bruxelles **18,** 1-38, 1845.

19. Montroll, E.W., **Social dynamics and the quantifying of social forces**, Proc. Nat. Acad. Sci. USA **75**, 4633-4637, 1978.

20. Jefferson, M. **The law of primate city**. Geographical Review, 29(2), 226-232, 1939.

21. Laherrere, J. & Sornette, D. **Stretched exponential distributions in nature and economy fat tails with characteristic scales**. European Physics Journal B, 2(4), 525-539, 1998.

22. R. Mansilla, E. Koppen, G. Cocho, P. Miramontes**, On the behavior of journal impact factor rank-order distribution**, Journal of Informetrics **1,** 155-160, 2007.

23. I. Popescu, **On a Zipf's Law Extension to Impact Factors**, Glottometrics **6,** 83-93, 2003.

24. I.A. Voloshynovska, **Characteristic Features of Rank-Probability Word Distribution in Scientific and Belletristic Literature**, Journal of Quantitative Linguistics **18,** 274-289, 2011.

25. C. Tsallis, M.P. de Albuquerque, **Are citations of scientific papers a case of nonextensivity?**, Eur. Phys. J. B **13,** 777-780, 2000.

26. E.W. Montroll, M.F. Shlesinger, **Maximum entropy formalism, fractals, scaling phenomena, and 1/f noise: a tale of tails**, J. Stat. Phys. 32, 209-230, 1983.

27. C. Rose, D. Murray, D. Smith, **Mathematical Statistics with Mathematica**, Springer, New York, p. 107, 2002.

28. Gompertz, R., **On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies**, Philos. Trans. R. Soc. London*,* **115**, 513-585, 1825.

29. Dimitrova, Zlatinka and Ausloos, Marcel. **Primacy analysis of the system of Bulgarian cities**, arXiv preprint: arXiv: 1309.0079, 2013

30. Nikolay K. Vitanov, Zlatinka I. Dimitrova, Marcel Ausloos, **City population sizes and power laws. The case of Bulgarian cities**, submitted

31. Marcel Ausloos and Roy Cerqueti, **Local and regional disparities of urban hagiotoponyms in Italy**, in preparation

32. D. Sornette, A. Johansen, and J.P. Bouchaud, **Stock Market Crashes, Precursors and Replicas**, J. Phys. I (France) 6, 167-175, 1996.

33. N. Vandewalle and M. Ausloos, **How the financial crash of October 1997 could have been predicted**, Eur. J. Phys. B 4, 139-141, 1998.

34. A. Johansen and D. Sornette, **The NASDAQ crash of April 2000: Yet another example of log-periodicity in a speculative bubble ending in a crash**, Eur. J. Phys. B 17, 319-328, 2000.

35. M. Ausloos, K. Ivanova, and N. Vandewalle, **Crashes: symptoms, diagnoses and remedies, in Empirical sciences of financial fluctuations. The advent of econophysics**, H. Takayasu, Ed. Springer Verlag, Berlin, pp. 62-76, 2002.

36. J. Kwapien and S. Drozdz, **Physical approach to complex systems**, Physics Reports 515, 115-226, 2012.

37. A. Johansen, D. Sornette, H. Wakita, U. Tsunogai, W.I. Newman, and H. Saleur, **Discrete Scaling in Earthquake Precursory Phenomena: Evidence in the Kobe Earthquake, Japan**, J. Phys. I (France) 6, 1391-1402, 1996