Size-frequency and rank-frequency relations, power laws and exponentials: a unified approach

Leo EGGHE and Ronald ROUSSEAU

(Limburgs University Center, Diepenbeek, B-3590 Belgium; University of Antwerp, IBW, Universiteitsplein 1, B-2610 Belgium)

Received November 1, 2002; revised December 5, 2002

Abstract Power laws, such as Zipf's law, and exponential relations, leading to straight lines in logarithmic or semi-logarithmic scales, are presented in a unified setting. It is shown that the class of size-frequency power laws is larger than the class of rank-frequency power laws. Their ubiquity in all fields of science is illustrated.

Keywords: power laws, Lotka's law, Zipf's law, exponentials, unification.

A handful of scientists produces most innovations in research and development; the ten most populous countries contain more than 50% of the total population in the world; a small group of words are spoken very often, many words are rarely used; small numbers of large firms co-exist with a large number of smaller companies[1]. All these examples illustrate the fact that productivity, in the widest possible interpretation of a relation between sources and items, is highly skewed. In recent times it has been shown that the distribution of in- and out-links on the Internet, the distribution of web site sizes (number of pages), and web site use, all are skewed phenomena[2~6].

Distributions such as the ones hinted at above can be described, at least in two ways: using a size-frequency and using a rank-frequency approach. Using size-frequencies means giving a functional relation, $f(n)$, $n = 1, 2, 3, \cdots$, between the number of sources and the number of items they produce. In the rank-frequency approach one first ranks the sources in decreasing order of production, and then finds a functional relation between the rank and the production of the source at that rank.

The classical expression for the size-frequency approach is a power law, often referred to as Lotka's law[7]:

$$f(n) = \frac{C}{n^{\alpha}}, \qquad (1)$$

where C and $\alpha > 0$ are constants. Although other functions have been proposed, the power law is the only one that is scale-free[8]. This means that if the measurement scale is changed (say n becomes constant $* n$) then, up to a constant, the same power law is obtained. The term "self-similarity" is often used in this context.

For rank-frequency distributions scientists usually prefer Zipf's law[9]. This relation is given as:

$$g(r) = \frac{D}{r^{\beta}}, \qquad (2)$$

where r denotes a rank. This relation is named after George K. Zipf, but goes back to Estoup and Auerbach in the beginning of the 20th century. Eq. (2) has been generalized by Mandelbrot[10], the inventor of fractals, leading to:

$$g(r) = \frac{E}{(1 + Fr)^{\beta}}. \qquad (3)$$

In informetrics, the field where scientists mathematically model regularities in the information sciences, both approaches are routinely considered, while in most other fields one of the two—usually the rank-frequency form—is preferred. Practical investigations usually confirm the existence of relations (2) or (3), at least as a general trend.

Studying the literature on rank-frequency distributions, the question came up as what would happen if the power function were replaced by a decreasing exponential:

$$g(r) = Ge^{-ar}, \quad G > 0, \ a > 0. \qquad (4)$$

It will be shown that the exponential form is just a limiting case of the power relation. More precisely, the exponential rank-frequency distribution corresponds to a size-frequency power relation with exponent $\alpha = 1$, while the other cases correspond to exponents strictly larger than 1 (exponents between 0

and 1 are also possible, but of no interest here).

1 A continuous setting

A manageable mathematical theory of rank and size frequencies can only be considered in a continuous framework. Consider, for instance, the defining relation between the functions f and g:

$$g^{-1}(n) = r(n) = \sum_{m=n}^{n_{max}} f(m). \quad (5)$$

The function g^{-1} is the inverse function of g, and n_{max} denotes the production of the most productive source (the source at rank 1). Many sums like the one in Eq. (5) cannot be evaluated (exactly, or even approximately), while it is in most cases feasible to evaluate the corresponding integrals. This is particularly true in the case at hand. Such a continuous theory has been developed a decade ago by Egghe[11], but is little known outside the field of informetrics. Notations and main results are briefly reviewed here.

The discrete functions f and g are replaced by continuous density functions φ and χ. The symbol φ denotes a density of items per source, while the function χ refers to a density of sources per item. These functions play a dual role. The function φ (the continuous analogue of the size-frequency function f) is defined as:

$$\varphi: [1, \rho] \to \mathbb{R}^+: x \to \varphi(x), \quad (6)$$

where ρ denotes the maximum density (corresponding to n_{max}, the production of the most productive source). The defining relation (5) becomes now:

$$\chi^{-1}(y) = r(y) = \int_y^{\rho} \varphi(x)\mathrm{d}x$$

or

$$\varphi(x) = -\frac{1}{\chi'(\chi^{-1}(x))}. \quad (7)$$

It can be shown that the following assertions (i) and (ii) are equivalent:

(i) $\varphi(x) = \frac{C}{x^{\alpha}}$, $x \in [1, \rho]$, $C > 0$, $\alpha > 1$ with C and α constants;

(ii) $\chi(y) = \frac{E}{(1+Fy)^{\beta}}$, $y \in [0, T]$, $E, F > 0$, $\beta > 0$ with E, F and β constants. Here T corresponds to the total number of sources, being equal to the largest value of y. From Eq. (7) it follows that

$$T = \int_1^{\rho} \varphi(x)\mathrm{d}x. \quad (8)$$

Furthermore, the exponents α and β are related as:

$$\beta = \frac{1}{\alpha - 1}. \quad (9)$$

Zipf's law (Eq. (2)) may be considered as a special case of Mandelbrot's (Eq. (3)), by the substitution of r by $r' = r + 1$ (discrete case). In the continuous case such a shift transforms the domain $[0, T]$ into $[1, T+1]$.

It can even be shown[11] that the above result is also valid for $0 < \alpha < 1$. Of course, Eq. (9) is not valid for $\alpha = 1$. This is studied next.

The relation between φ and χ for $\alpha = 1$ is given in the following theorem.

Theorem. If $\alpha = 1$ the following equivalence holds:

(i) $\varphi(x) = \frac{C}{x}$, $x \in [1, \rho]$, $C > 0 \Leftrightarrow$ (ii) $\chi(y) = Ge^{-ay}$, $y \in [0, T]$, $G > 1$, $0 < a < +\infty$ with C, G and α constants. This shows that a size-frequency power law with $\alpha = 1$ is equivalent with an exponentially decreasing rank-frequency function.

Proof. First part: (i)$\Rightarrow$(ii). By Eq. (7) it is clear that

$$\chi^{-1}(t) = r(t)\int_t^{\rho} \frac{C}{x}\mathrm{d}x = C\ln\left(\frac{\rho}{t}\right).$$

Hence: $t = \chi(y) = \rho e^{-y/C}$, where $\rho = G > 1$, and $0 < 1/C = a < +\infty$.

Second part: (ii)$\Rightarrow$(i). Since $\chi(y) = Ge^{-ay}$ it follows that

$$\chi'(y) = -aGe^{-ay}$$

and

$$t = \chi^{-1}(x) = -\frac{\ln(x/G)}{a}.$$

Finally, Eq. (6) leads to:

$$\varphi(x) = \frac{1}{aGe^{\ln(x/G)}} = \frac{1}{ax}.$$

This is the required relation with $C = 1/a > 0$.

This result demonstrates that the class of size-frequency power laws is larger than the class of rank-frequency power laws. The fact that Lotka's power law with $\alpha = 1$ leads to an exponentially decreasing rank-frequency relation has been observed before[12], but no special attention has been paid to it.

Returning to the afore-mentioned examples, it is concluded that Lotka's size-frequency power law is even more ubiquitous than previously thought. Examples of exponential rank-frequency relations are in-

deed also examples of power law size-frequency relations (with $\alpha = 1$).

2 More exponentials

A rank-frequency power law leads to a straight line when represented using double-logarithmic scales. Similarly, an exponential rank-frequency relation, when represented using semi-logarithmic scales with a logarithmic scale on the vertical axis: (a so-called linear-log plot) leads to a straight line. Clearly, one may wonder what happens if a straight line is observed using semi-logarithmic scales, this time with the logarithmic scale on the horizontal axis (a log-linear plot). More interestingly, it is studied if such a relation has been observed in reality, and how it can be related to a power law.

A straight line in semi-logarithmic scales, logarithm on the horizontal axis, can be expressed as:

$$\chi(y) = -a\ln(y) + b, \quad a > 0, b > 0.$$

It is, moreover, assumed that $y \in [1, T+1]$. This corresponds to a shifted distribution, as explained before. Under these circumstances, $\chi'(y) = -\frac{a}{y}$ and $\chi^{-1}(t) = e^{\frac{b-t}{a}}$. Consequently, the corresponding size-frequency law is:

$$\varphi(x) = \frac{1}{a} e^{\frac{b-x}{a}}$$

where b can be interpreted as the production of the source with highest production. Such an exponential size-frequency relation is, of course, different from a power law. Yet, it is related to it in the sense that it has been observed in circumstances where power laws are expected, but which are artificial or more restricted (made homogeneous). Seglen[13], for instance, found such a relation when studying citations to one particular journal (and not to a whole field, or another larger group). Amaral et al.[14] showed figures of different small-world networks. The connectivity of the electric power grid of Southern California clearly has an exponential size-frequency distribution, and so does the incoming and outgoing neuronal links in the worm *Caenorhabditis elegans*. Huber[15] and Huber and Wagner-Döbler[16] found exponential size-frequency relations for the production of authors with the same career duration (another case of a restricted circumstance) and for career duration themselves. Finally, it is well-known that also the tail of the link distribution in a random network decreases exponentially[17]. This proves that such special cases do happen in real situations.

3 Conclusion

It is shown how a unified approach, including size-frequency and rank-frequency distributions, relates power laws and exponential relations occurring in all fields of science. The ubiquity of these relations in all fields of science is illustrated.

References

1 Axtell, R. L. Zipf distribution of US firm sizes. Science, 2001, 293: 1818.
2 Rousseau, R. Sitations: an exploratory study, Cybermetrics [e-journal], 1997, 1(1) http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html
3 Barabási, A. L. et al. Emergence of scaling in random networks. Science, 1999, 286: 509.
4 Hayes, B. Graph theory in practice: Part II, American Scientist, 2000, 88(2): 104.
5 Adamic, L. A. et al. The Web's hidden order, Communications of the ACM, 2001, 44(9): 55.
6 Adamic, L. A. et al. Zipf's law and the Internet. Glottometrics, 2002, 3: 143.
7 Lotka, A. The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences, 1926, 16: 317.
8 Roberts, F. S. Measurement Theory. Reading: Addison-Wesley, 1979.
9 Zipf, G. K. Human Behavior and the Principle of Least Effort. Cambridge: Addison-Wesley, 1949.
10 Mandelbrot, B. Structure formelle des texts et communication, Word (in French) 1954, 10:1.
11 Egghe, L. The duality of informetric systems with applications to the empirical laws. Journal of Information Science, 1990, 16: 17.
12 Rousseau, R. Lotka's law and its Leimkuhler representation. Library Science with a Slant to Documentation, 1988, 25: 150.
13 Seglen, P. The skewness of science. Journal of the American Society for Information Science, 1992, 4: 628.
14 Amaral, L. A. N. et al. Classes of small-world networks. In: Proceedings of the National Academy of Science USA, 2000, 97: 11149.
15 Huber, J. C. The underlying process generating Lotka's law and the statistics of exceedances. Information Processing and Management, 1998, 34: 471.
16 Huber, J. C. et al. Scientific production: a statistical analysis of authors in mathematical logic. Scientometrics, 2001, 50: 323.
17 Albert, R. et al. Error and attack tolerance of complex networks. Nature, 2000, 406: 378.