Malaysian Journal of Library & Information Science, Vol. 17, no. 3, December, 2012: 1-5

Theoretical evidence for empirical findings of A. Pulgarin on Lotka's law

L. Egghe
Universiteit Hasselt, Campus Diepenbeek, Agoralaan,
B – 3590 Diepenbeek, BELGIUM[1]
Universiteit Antwerpen, Stadscampus, Venusstraat 35,
B – 2000 Antwerpen, BELGIUM
e-mail: leo.egghe@uhasselt.be

ABSTRACT

*In [*A. Pulgarin. 2012. Dependence of Lotka's law parameters on the scientific area*. Malaysian Journal of Library and Information Science*, *17(1): 41-50], the author finds negative correlations between the average number of papers per author and the Lotka exponent and the Lotka constant as well as positive correlations between the latter two parameters. He also finds lower values for the latter two parameters in fields or countries with highly concentrated productions (e.g. where there is heavy growth). In the present paper these findings are proved mathematically, based on earlier results on Lotka's law proved by this author.*

Keywords: Lotka's law; Zipf's law; Scientometrics

INTRODUCTION

Lotka's law (Lotka (1926)) states that the fraction of authors $f(n)$ with $n = 1, 2, 3, \ldots$ publications equals

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

where $C > 0$ and $\alpha > 1$ are parameters. C is called the Lotka constant and α is called the Lotka exponent. In its continuous form (for $n \geq 1$), $f(n)$ is the density of authors with publication density *n*.

In an extensive experiment (covering several disciplines, countries and time periods), Pulgarin (2012) finds several regularities between the parameters C, α and μ (being the average number of papers per author). He finds a positive correlation between C and α and a negative correlation between μ and C and between μ and α. Furthermore he finds that in dynamic fields or countries (where there is a heavy growth) there is a high concentration in articles, being spread out over the authors, leading to lower values of the Lotka parameters C and α.

[1] Permanent address

In this paper we will give mathematical explanations of these experimental findings based on earlier results of this author and extensively described in Egghe (2005). All these experimental findings are mathematically confirmed. The contradiction between Pulgarin (2012) and Wagner-Döbler and Berg (1995) is explained as an error in the latter paper.

Explanations

Let us take (1) as a continuous distribution, i.e. for which

$$\int_1^\infty f(n)dn = \int_1^\infty \frac{C}{n^\alpha}dn = 1 \tag{2}$$

Hence

$$C = \frac{1}{\int_1^\infty \frac{1}{n^\alpha}dn} \tag{3}$$

When α increases, n^α increases (as *n* is at least one), hence $\frac{1}{n^\alpha}$ decreases and the integral also decreases. So, formula (3) already shows that the Lotka constant C is an increasing function of *α*, which explains the experimentally found positive correlation between C and *α* (there denoted *c(t)* and *n(t)* respectively) in Pulgarin (2012). Indeed, as shown in Egghe and Rousseau (1996) an increasing function leads to a positive correlation.

Note that in 87 out of the 90 datasets in Pulgarin yield a Lotka exponent $\alpha > 2$ In case $\alpha > 2$ we have that

$$T = \int_1^\infty f(n)dn = \frac{C}{\alpha - 1} \tag{4}$$

is the total number of authors and that

$$A = \int_1^\infty nf(n)dn = \frac{C}{\alpha - 2} \tag{5}$$

is the total number of papers. We hence have that *μ*, the average number of papers per author, equals

$$\mu = \frac{A}{T} = \frac{\alpha - 1}{\alpha - 2} \tag{6}$$

(see also Egghe (2005)). This shows that *μ* is a decreasing function of *α*, which explains the empirically found negative correlation between *μ* and *α* in Pulgarin (2012). Since we showed already that C is an increasing function (3) of *α*, we hence also have that *μ* is a decreasing function of C. Hence this also explains the empirically found negative correlation between *μ* and C in Pulgarin (2012).

In order to be able to further explain experimental results in Pulgarin (2012), we have to introduce Zipf's law: let $g(r)$ denote the number of papers of the author on rank $r = 1, 2, ..., T$ (T= total number of authors). Then

$$g(r) = \frac{B}{r^\beta} \tag{7}$$

where B, $\beta > 0$ are parameters. In the continuous setting, $g(r)$ denote the density of papers on rank - density $r \in [0,T]$. In this setting, Lotka's law (for $\alpha > 1$) and Zipf's law are equivalent and the relation between their exponents is

$$\beta = \frac{1}{\alpha - 1} \tag{8}$$

, see Egghe (2005), Exercise II.2.2.6 or Egghe and Rousseau (2006), Appendix, where a proof of this equivalence is given. Note that (8) is a decreasing function: β is a decreasing function α. This fact has produced confusion in some informetric publications (Wagner-Döbler and Berg (1995), Yoshikane and Kageura (2004)). We will explain this further on which will also explain why Pulgarin (2012) contradicts Wagner-Döbler and Berg (1995). We will show that Pulgarin (2012) gives the correct interpretation.

The continuous form of the Lorenz curve is needed to explain other experimental results in Pulgarin (2012). We explain this now. Denoting by $h(x)$, $x \in [1, x_m]$

$$h(x) = \frac{K}{x^{\gamma}} \tag{9}$$

any of the functions (1) or (7), the Lorenz curve of the function $h(x)$ is given by the following set of points for $x \in [1, x_m]$

$$\left(\frac{x-1}{x_m - 1}, \frac{\int_1^x h(x')\,dx'}{\int_1^{x_m} h(x')\,dx'} \right) \tag{10}$$

i.e. the cumulative fraction of the abscissae versus the cumulative fraction of the ordinates for the function $h(x)$. In other words, putting

$$y = \frac{x-1}{x_m - 1} \tag{11}$$

(hence $y \in [0,1]$ and also: $x = y(x_m - 1) + 1$), the Lorenz curve $L(h)$ of *h* is the function

$$L(h)(y) = \frac{\int_1^{y(x_m-1)+1} h(x')\,dx'}{\int_1^{x_m} h(x')\,dx'} \tag{12}$$

(see also Egghe (2005); p.196 and further).

The Lorenz curve was introduced in Lorenz (1905) in econometrics in order to be able to calculate the inequality (or concentration) of the function $h(x)$. The Lorenz curve (12) is a concavely increasing function connecting (0, 0) and (1, 1) and we have that, the higher the Lorenz curve $L(h)$, the higher the concentration (inequality) of the function $h(x)$. Denoting by $L(f)$ and $L(g)$ the Lorenz curves of the Lotka function (1) and the Zipf function (7) respectively we have proved in Egghe (2005), p.204-205, the following key result:

Theorem (Egghe 2005) : Let f and g be as above. Then the following assertions are equivalent:

(i) $L(f)$ strictly increases in α

(ii) $L(g)$ strictly increases in β

(iii) $L(g)$ strictly decreases in α

Note that the equivalence of (ii) and (iii) follows from the decreasing function (8).

In emerging fields or countries Pulgarin (2012) finds author distributions which are very concentrated. Hence this refers to the Zipfian function $g(r)$. Increasing concentration of $g(r)$ means increasing Lorenz curves $L(g)$. According to the above theorem we have decreasing α-values and, because of (3), hence also decreasing C-values. This explains the findings in Pulgarin (2012) who finds low values of α and C for such concentrated situations.

That these findings contradict the ones in Wagner-Döbler and Berg (1995) is due to the fact that this paper refers to the concentration of the Lotka function $f(n)$, hence referring to $L(f)$. Because of the above theorem, the conclusions must be opposite to the ones of Pulgarin (2012) which was indeed the case. A similar confusion appeared in Yoshikane and Kageura (2004).

CONCLUSIONS

In this paper, based on existing results on Lotka's law (Egghe 2005), we could explain several experimental findings in Pulgarin (2012):

a) The average number μ of papers per author correlates negatively with the Lotka parameters C and α,
b) The Lotka parameters C and α correlate positively,
c) In case of high concentration of authors (e.g. in emerging fields or countries), the Lotka parameters C and α are small.

It is the hope that this paper contributed to the theoretical explanation of these findings and that it solves the confusion among some authors that arises by considering Lotka's law instead of Zipf's law or vice-versa.

REFERENCES

Egghe, Leo. 2005. *Power laws in the information production process.* Oxford: Elsevier.

Egghe, L. and R. Rousseau. 1996. Average and global impact of a set of journals. *Scientometrics*, Vol. 36, no.1: 97-107.

Egghe, L. and R. Rousseau. 2006. An informetric model for the Hirsch-index. *Scientometrics*, Vol. 69, no.1: 121-129.

Lotka, A.J.. 1926. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, Vol. 16, no. 12: 317-324.

Lorenz, M.O. 1905. Methods of measuring concentration of wealth. *Journal of the American Statistical Association*, Vol. 9: 209-219.

Pulgarin, A. 2012. Dependence of Lotka's law parameters on the scientific area. *Malaysian Journal of Library and Information Science,* Vol. 17, no.1:41-50.

Wagner-Döbler, L. and J. Berg. 1995. The dependence of Lotka's law on the selection of time periods in the development of scientific areas and authors. *Journal of Documentation*, Vol. 51, no.1: 28-43.

Yoshikame, F. and K. Kageura. 2004. Comparative analysis of coauthorship networks of different domains: the growth and change of networks. *Scientometrics*, Vol. 60, no.3: 433-444.