INFORMETRICS 89/90
L. Egghe and R. Rousseau (Editors)

THE CUMULATIVE ADVANTAGE FUNCTION. A MATHEMATICAL FORMULATION BASED ON CONDITIONAL EXPECTATIONS AND ITS APPLICATION TO SCIENTOMETRIC DISTRIBUTIONS

Wolfgang GLAENZEL, Andras SCHUBERT

Department of Scientometrics, Library of the Hungarian Academy of Sciences, H-1361, P.O. Box 7, Budapest (Hungary)

Abstract

Cumulative advantage principle is a specific law underlying several social, particularly, bibliometric and scientometric processes. This phenomenon was described by single- and multiple-urn models (Price (1976), Tague (1981)). A theoretical model for cumulative advantage growth was developed by Schubert and Glaenzel (1984). This paper presents an exact measure of the cumulative advantage effect based on conditional expectations. For a given bibliometric random variable X (e.g. publication activity, citation rate) the cumulative advantage function is defined as $\mu(k) = E((X-k)|(X-k) \geq 0)/E(X)$. The 'extent of advantage' is studied on the basis of limit properties of this function. The behavior of $\mu(k)$ is discussed for the urn-model distributions, particularly for its most prominent representants, the negative-binomial and the Waring distribution. The discussion is illustrated by several examples from bibliometric distributions.

1. *INTRODUCTION*

The phenomenon of cumulative advantage is intimately connected with social processes. It is in effect whenever a new event is influenced by previous successes or failures. Therefore the principle is also called success-breeds-success-phenomenon. It was shown (e.g. Price (1976), Tague (1981)) that the cumulative advantage principle (c.a.p.) underlies several bibliometric/scientometric phenomena. Other bibliometric phenomena may just as well be independent of any success and failure influences. Given a set of empirical data it is not always easy to decide whether c.a.p. is underlying or not. Therefore an effective measure and, first of all, a definition of the cumulative advantage principle is needed. A verbal description was given by de Solla Price (1976) :

"A paper which has been cited many times is more likely to be cited again than one which has been little cited. An author of many papers is more likely to publish again than one who has been less profilic. A journal which has been frequently consulted for some purpose is more likely to be turned to again than one of previously infrequent use. Words become common or remain rare. A millionaire gets extra income faster and easier than a beggar."

The present paper attempts to give an exact mathematical description and a quantitative measure of this phenomenon. The mathematical results are applied to bibliometric distributions of several types.

2. THE STOCHASTIC MODEL

In order to visualize the background and mechanism of the cumulative advantage principle a stochastic process introduced by Schubert & Glaenzel (1984) is used. We recall the model in brief :
Consider an infinite array of units indexed in succession by the non-negative numbers. The content of the i-th unit is denoted by x_i, the (finite) content of all units by x. Then the fraction $y_i = x_i/x$ $(i \geq 0)$ expresses the (classical) probability with which an element is contained by the i-th unit. The stochastic process is formed by the change of the content of the units. The change is postulated to obey the following three rules :

(1) Substance may enter the system from the external environment through the 0-th unit at a given rate.
(2) Substance may be transferred from any unit to the adjoining ones at a given rate.
(3) Substance may leak out from any unit into the external environment at a given rate.

It is clear that the content of the units has a specific distribution at any time. We assume that at the beginning only the 0-th unit contains elements, i.e. the entire population is concentrated in the 0-th unit. We further assume that the process has a non-degenerated limit distribution as time tends to infinity. Fig.1 shows the scheme of substance flow of this process.

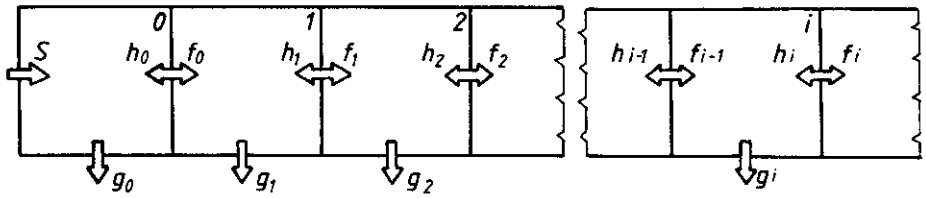

Fig.1 : Scheme of substance flow in the stochastical cumulative advantage model

Cumulative advantage means that an element being in a unit with a high index is more likely to reach a unit with a still higher index than an element of a unit with a low index. The step from the i-th to the (i+1)-th unit can be considered as a success, a step in the reverse direction as a failure. A cumulation of successes and failures during the time elapsed is determined by the transfer rates and results in a specific limit distribution. Since, in general, empirical data give information about the stationary limit distribution, it seems reasonable to define a cumulative advantage measure on this limit distribution.

3. THE CUMULATIVE ADVANTAGE FUNCTION

The c.a.p. is partially reflected by the expected remaining life function originally defined for renewal processes (e.g. Kotz & Shanbagh (1980)). For a non-negative integer-valued random variable X, it is defined as $s(k) = E(X-k \mid X \geq k)$, $k = 0,1,2,\ldots$, provided the expectation is finite. In the frame of the stochastic process of the preceding section s(k) is the expected index of the unit to be reached by an element provided the k-th unit had already been reached. A cumulative advantage then corresponds to an increasing function s(k). Unfortunately, for our purpose the expected remaining life function must be rejected because this function depends on the expectation. Thus it measures not only the "acquired" advantage rate but

the "innate" advantage, too.

In order to overcome this shortcoming we define the cumulative advantage function (c.a.f.) as

$$\mu(k) = \frac{E((X-k)|(X-k) \geq 0)}{E(X)}$$

This function obviously meets our requirements. The analysis of the c.a.f. should cover (1) the monotonicity and (2) the limit at infinity.
Ad (1). Two cases are of particular interest in our present analysis : if μ is monotone or if μ can be splitted into two monotone parts. Other special patterns (e.g., periodical behavior) could be the topic of a separate study.
Ad (2). Concerning the limit, 5 different types can be distinguished. Put
$c = \lim_{k\to\infty} \mu(k)$

1. $c = \infty$ (cumulative advantage principle)
2. $1 < c < \infty$ (limited influence of advantage)
3. $c = 1$ (no advantage at all)
4. $0 < c < 1$ (limited influence of disadvantage)
5. $c = 0$ (cumulative "disadvantage" principle)

Note that $\mu(0) = 1$ and $\mu(k) \geq 0$ for all k. Before applying this function to any particular distribution it would be worthwhile having a look at the relationship between the cumulative advantage principle and the tail properties of distributions.

Theorem 1 :

Let p_k denote the probability $P(X=k)$ and put $p_{k+1}/p_k = q_k$, $\lim_{k\to\infty} q_k = q$.

Assume that c is the same as above. Then the following equivalent statements hold :

(1) $c = 0 \iff q = 1$
(2) $0 < c < \infty \iff 0 < q < 1$
(3) $c = 0 \iff q = 0$.

Proof :

The proof is based on characterization theorems by Gupta (1975) and Glaenzel & al. (1984). According to these results the distribution p_k is uniquely determined both by the expected remaining life function and by the c.a.f. and E(X). Thus we have :

$$p_k = \prod_{i=0}^{k-1} \frac{\mu_i}{\mu_{i+1} + d} \frac{\mu_{k+1} - \mu_k + d}{\mu_{k+1} + d} ,$$

where $d = 1/E(X)$. Hence

$$q_k = \frac{\mu_{k+1}}{\mu_{k+1} + d} \frac{\mu_{k+2}/\mu_{k+1} - 1 + d/\mu_{k+1}}{\mu_{k+1}/\mu_k - 1 + d/\mu_k}$$

is obtained. The right hand factor of the above equation always tends to 1, independently of the actual limit $c = \mu_\infty$. Therefore the following

approximation can be studied instead :

$$q_k \approx \mu_{k+1}/(\mu_{k+1} + d)$$

Hence the statement can directly be derived.

Now we show the close relationship between the tail of a distribution and the limit value of the c.a.f.
A distribution is said to have a proper tail, if it asymptotically obeys "Zipf's Law", e.g., if

$$\lim_{k\to\infty} \Big(\sum_{i=k}^{\infty} p_i \Big) k^{\alpha} = \text{const}$$

for some real $\alpha > 0$ (see Glaenzel & Schubert (1988)). This already implies the convergence $p_{k+1}/p_k \approx (1 + 1/k)^{\alpha} \to 1$. This consideration and the Theorem 1 leads to the following important result :

Theorem 2 :

Let X be a non-negative integer-valued random variable the distribution of which has a proper tail with the characteristic exponent $\alpha > 1$. Then the cumulative advantage principle underlies the distribution of X, i.e., $\mu(k) \to \infty$ as $k \to \infty$.

4. THE URN MODEL DISTRIBUTIONS

The connection between the cumulative advantage principle and some representatives of the urn model distributions has already been shown and discussed in the papers by Price (1976) and Tague (1981). In the following the c.a.f. introduced in the preceding section will be applied to the distributions of the classical Polya-Eggenberger urn model. First of all we recall how this single-urn model works. Let an urn contain a certain number of black and white balls. White usually means success, black means failure. A ball is drawn from the urn at random. Together with the drawn ball a certain fixed number of white or black balls is replaced into the urn, according as the color of the drawn ball was white or black, respectively. The number of added balls can be positive, zero or negative. The process of drawing and replacing balls is continued according to the chosen model. Two models, a "finite" (1) and an "infinite" (2) one are used.

(1) The process is stopped after the n-th ball is drawn.
(2) The process is stopped if the n-th black ball (failure) is drawn.

Random events are the number of the drawn white balls (successes). The built-in cumulative advantage-disadvantage phenomenon is obvious. The proof of the following statement can be found in the literature (e.g. Johnson & Kotz (1977)).

Proposition :

The distribution of a non-negative integer-valued random variable X can be obtained from the urn model, if

$$P(X=k) = \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+\tau)}{\Gamma(\alpha)\Gamma(\alpha+\beta+\tau)} \frac{\beta\tau}{(\alpha+\beta+\tau)} \cdots \frac{(\beta+k-1)(\tau+k-1)}{(\alpha+\beta+\tau+k-1)k} \quad ,$$

where α, β and τ are parameters such that p_k is a probability for every $k \geq 0$.

The connection with the cumulative advantage principle can be demonstrated by two examples. Model (2) is assumed, i.e., balls are drawn until a certain number of black balls occur.

1. Waring distribution : the number of added balls after each drawing is positive. Procedure is stopped when the first black ball is drawn. Thus the number of drawn white balls is influenced only by successes but not by failures. This is a pure cumulative advantage distribution.

2. Negative-binomial distribution. No balls are added when replacing the drawn one. Procedure is stopped after the n-th black ball was drawn. No cumulative advantage effect.

Now we have a look at the c.a.f. of these distributions. An important property of the c.a.f. of urn model distribution is reflected by Theorem 2 which is a result of a theorem by Glaenzel & Schubert (1985) :

Theorem 3 :

The cumulative advantage function of a distribution of the Polya-Eggenberger urn model is always asymptotically linear, provided the expectation is finite, in particular :

$$\mu(k) \approx k/\beta\tau + 1 - \alpha\,(1 - 1/\beta)(1 - 1/\tau)/(\alpha + 1), \quad k \gg 1 \ .$$

It is interesting to note that the c.a.f. of a Waring distribution is linear and completely independent of the characteristic tail parameter α : $\mu(k) = k/N + 1$. The c.a.f. of a negative-binomial distribution has the asymptotic equation $\mu(k) \approx 1/N$, i.e., the limit value does not depend on the parameter q. Based of a comparison with the classification of Section 3 we can state that a negative binomial distribution reflects a limited influence of advantage if $N < 1$, a limited influence of disadvantage if $N > 1$ and no advantage, if $N = 1$ (geometric distribution). A cumulative "disadvantage" principle underlies the Poisson and all finite urn model distributions ($\mu(k) \approx 0$, if $k \gg 1$). Figure 2 illustrates the behavior of the c.a.f. by means of four urn model distributions with different parameters.

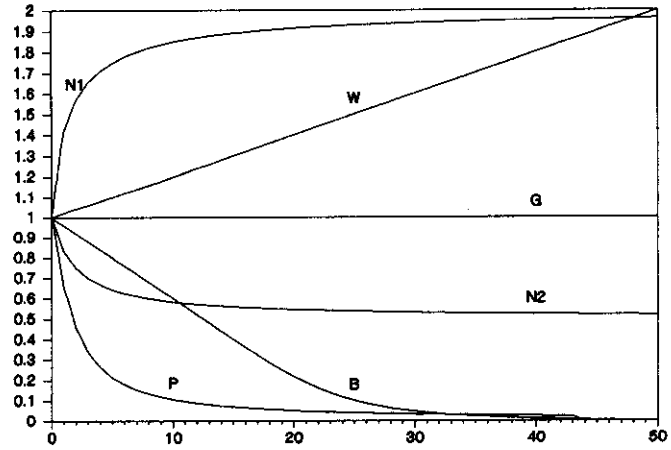

Fig.2 : c.a.f. of some representatives of the urn model distributions : W = Waring with N = 50, α optional; N1 = negative binomial with N = .5, q = .5; N2 = negative binomial with N = 2, q = .5, P = Poisson with λ = 2 and B = binomial with n = 50, p = .5.

5. APPLICATION TO BIBLIOMETRIC DISTRIBUTIONS

In this section we illustrate on bibliometric examples how a cumulative advantage effect can be detected and interpreted on the basis of the c.a.f. Four basic types of scientometric distributions were chosen for the analysis. Two of them are related to the reference/citation process, the other two distributions are connected with the publication/authorship process. The first pair of distributions (a citation rate and a reference distribution) are taken from the SCI database of the Institute of Scientific Information (Philadelphia, USA). The data are restricted to papers (of the 5-year time period 1981-85) with at least one Hungarian co-author. Citation rates has been counted for the same time period, references were considered without any restrictions. In order to obtain a homogeneous population, only research papers were taken into account. Thus, among others, the items "review or bibliography" were deliberately omitted because they would cause an artificially long tail of the reference distributions. Citation rates are, of course, expected to reflect a cumulative advantage effect. Fig.3 shows the diagram, c.a.f., $\mu(k)$ vs. citations k for $k \leq 50$. As expected, an increasing c.a.f. is obtained. The points $(k,\mu(k))$ form an almost perfect straight line. Thus the underlying cumulative advantage principle is unambiguous.

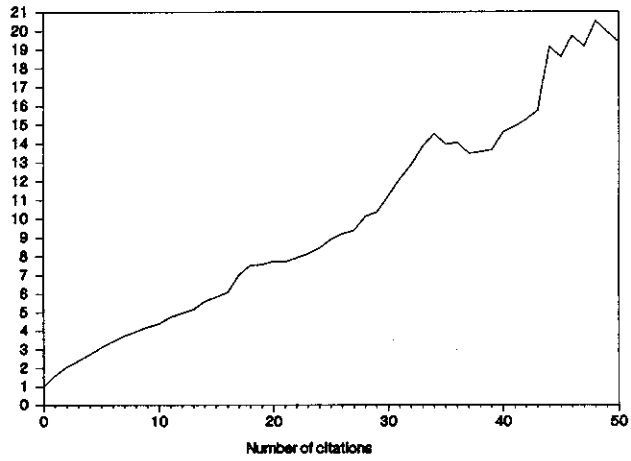

Fig.3 : c.a.f. of the citation rate distribution of Hungarian papers published in 1981-85

Reference data are not expected to reflect any effect of cumulative advantage. The decreasing c.a.f. at the beginning seems to confirm this assumption (see Fig.4). The behavior of the function for greater k ($k > 30$), however, contradicts the assumption. This phenomenon suggests the following explanation : The reference distribution is influenced by two tendencies. One relates to papers with only a few references. Only as many related papers as absolutely necessary and relevant are cited. When the most important references are included, further, less relevant ones are "repulsed". Thus in the beginning of the distribution a disadvantage is observed. The second tendency can be considered as some kind of "chain reaction". In this case, papers aim at completeness. Each cited paper uncovers a whole set of further references and some of them may be added to the reference list. The latter "2nd generation" references may attract a 3rd generation and so on. This leads to a kind of cumulative advantage. The distribution can be found in Table 1.

The data of the second pair of samples are taken from the data base MEDLARS. All papers concerning the research of the cancer medicament "Endoxan" were

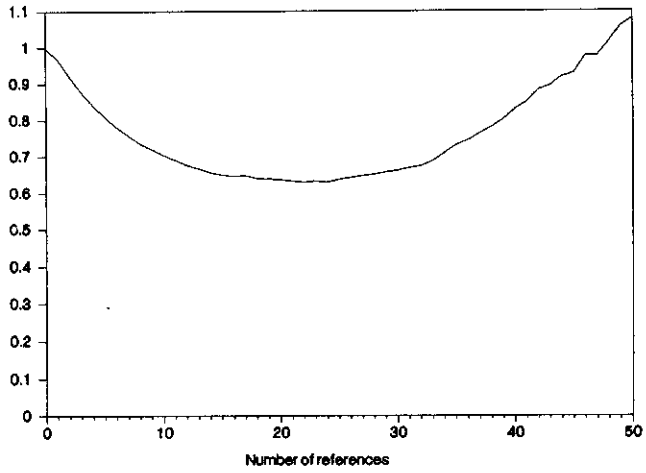

Fig.4 : c.a.f. of the reference distribution of Hungarian papers published in 1981-85

selected (for the 5-year period 1982-1986). The first sample, publication activity data, is a classical example for the cumulative advantage principle. The topic of the papers is very specific, therefore the author population is rather homogeneous. The second sample is the distribution of the number of authors of the same set of papers. The frequency distributions of both samples are presented in Table 2. The cumulative advantage function can be found in Fig.5. Although neither distribution has a proper tail, the c.a.f. clearly shows a cumulative advantage for the publication activity distribution and a disadvantage for the authorship distribution. In the case the productivity this phenomenon may be caused by the shortness of the time period (a relative long time is needed for the observations of the patients). The lacking tail of the author distribution can be explained by the fact that generally small groups of scientists are involved in this research, thus the number of potential authors is very limited. On the other hand, the distribution

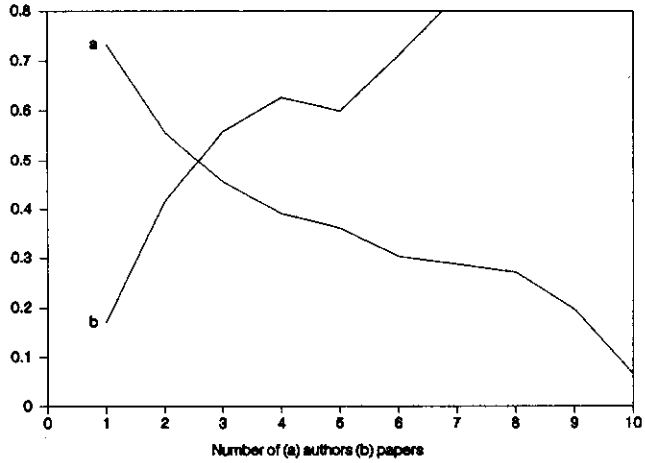

Fig.5 : c.a.f. of authorship (a) and publication activity (b) of 837 "Endoxan" papers and 2604 authors (1982-86)

Table 1 : The absolute frequency distribution of citation rates (X) and references of (Y) 10256 Hungarian papers published in 1981-85

k	X_k	Y_k
0	5521	232
1	1850	56
2	857	129
3	562	189
4	368	269
5	245	299
6	173	338
7	118	365
8	94	421
9	71	401
10	69	417
11	44	392
12	39	411
13	40	393
14	25	406
15	22	399
16	31	409
17	17	323
18	8	342
19	9	300
20	6	284
21	8	259
22	7	271
23	7	221
24	7	247
25	5	211
26	4	190
27	6	172
28	3	163
29	5	145
30	4	135
31	3	121
32	3	126
33	2	120
34	0	108
35	1	84
36	0	81
37	1	72
38	1	66
39	2	63
40	1	52
41	1	53
42	1	37
43	3	39
44	0	29
45	1	41
46	0	20
47	1	31
48	0	28
49	0	20
50	1	22
> 50	9	254

Table 2 : The absolute frequency distribution of publication activity (X) and authorship (Y) of 837 "Endoxan" papers published in 1982-86 (2604 authors)

k	X_k	Y_k
1	2247	91
2	250	171
3	66	177
4	23	150
5	11	91
6	4	71
7	2	40
8	0	19
9	0	11
10	1	13
11	0	2
12	0	1

is not extremely skewed, because most papers of this topic are team work. Single author papers are not too frequently observed. If we compare the behavior of the latter two distributions with the theoretical considerations of Section 4, we can claim, that the productivity distribution may probably be approximated by a Waring distribution, while a negative binomial distribution may give a good approximation for the authorship distribution.

REFERENCES

Glaenzel, W., Telcs, A., Schubert, A., Characterization by truncated moments and its application to Pearson-type distributions. Z. Wahrsch. Verw. Gebiete 66 (2) (1984), 173-183.

Gupta, R.C., On the characterization of distributions by conditional expectations. Commun. in Statist. 4 (1975), 99-103.

Johnson, N.L., Kotz, S., Urn models and their applications. John Wiley & Sons, New York-London-Sidney-Toronto, 1977.

Kotz, S., Shanbhag, D.N., Some new approaches to probability distributions. Advan. Appl. Probab. 12 (1980), 903-921.

Price, D. de Solla, A general theory of bibliometric and other cumulative advantage processes. JASIS 27 (5) (1976), 292-305.

Schubert, A., Glaenzel, W., A dynamic look at a class of skew distributions. A model with scientometric applications. Scientometrics 6 (3) (1984), 149-167.

Tague, J.M., The success-breeds-success phenomenon and bibliometric processes. JASIS 32 (4) (1981), 280-286.