JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE □
27 (5-6): 292-306 1976

A General Theory of Bibliometric and Other Cumulative Advantage Processes*

A Cumulative Advantage Distribution is proposed which models statistically the situation in which success breeds success. It differs from the Negative Binomial Distribution in that lack of success, being a non-event, is not punished by increased chance of failure. It is shown that such a stochastic law is governed by the Beta Function, containing only one free parameter, and this is approximated by a skew or hyperbolic distribution of the type that is widespread in bibliometrics and diverse social science phenomena. In particular, this is shown to be an appropriate underlying probabilistic theory for the Bradford Law, the Lotka Law, the Pareto and Zipf Distributions, and for all the empirical results of citation frequency analysis. As side results one may derive also the obsolescence factor for literature use. The Beta Function is peculiarly elegant for these manifold purposes because it yields both the actual and the cumulative distributions in simple form, and contains a limiting case of an inverse square law to which many empirical distributions conform.

Derek de Solla Price
Department of History of Science and Medicine
Yale University
New Haven, CT 06520

• Introduction

It is common in bibliometric matters and in many diverse social phenomena, that success seems to breed success. 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 prolific. 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.

In statistics, such a process is commonly described by a skew or hyperbolic distribution function of the type that has been characterized by Simon (*1*) and correctly shown by him to be given by the Beta Function (this is not the "Beta Density" which is discussed by Feller (*2*) II:49 and includes a dependent variable) rather than the more commonly used "contagious" distributions, for example the negative binomial (*3,4,5*) or its limiting form, Fisher's logarithmic series distribution.

Although the relation between such distributions, the stochastic processes which lead to them and the Urn models from which they may be derived, seem well known and go back more than 50 years to Yule's (*6*) first use of such probability models to explain the distribution of biological genera and species, it does not seem that the full elegance of the Beta Function is widely appreciated; certainly not in the context of bibliometrics though good theoretical treatments have been given by Hill (*7,8*), Crowley (*9*), Hill and Woodroofe (*10,11*) and Sichel (*12*), all discussing the general varieties of contagious process that yields hyperbolic distributions. As we shall show, this distribution which we propose to call the *Cumulative Advantage Distribution* (*CAD*), [after the pioneering sociological work in which Cole and Cole (*13*) following Merton discuss accumulative advantage in social stratification of the scientific community] can be derived either from a modification of the Polya Urn model, or as a stochastic birth process. It provides a

*This research was funded by NSF grant SOC73-05428.

sound conceptual basis for such empirical laws as the Lotka Distribution for Scientific Productivity, the Bradford Law for Journal Use, the Pareto Law of Income Distribution, and the Zipf Law for Literary Word Frequencies. It is therefore an underlying probability mechanism of widespread application and versatility throughout the social sciences. Though I cannot hope to deal adequately with all its mathematical aspects, this account should make the general theory more available and thereby remove much of the present restriction to empirical laws whose functions is useful but not fundamental.

• The Urn Model

The idea to use urn models to describe statistical after-effects or "contagion" seems to be due to Polya (see Feller (*2*) I:119, n.1). In general, the model supposes that fate has in storage an urn containing red and black balls; at regular intervals a ball is drawn at random, a red ball signifying a "success" and a black ball a "failure." If the composition of the urn remained fixed, the chances of success and failure would not vary, but if at each drawing the composition is *changed* by some rule, the *chances* will change as an after-effect of the previous history. A more general rule would be that after each drawing the ball is replaced and c balls of the color drawn and d balls of the opposite color are added before the next drawing takes place. For the Polya Urn Scheme, which can be shown to lead to a negative binomial distribution, $c > 0$, $d = 0$, so that at each drawing the number of balls of the color drawn *increases* while that of the other color remains unchanged. Thus, each occurrence of a red or of a black increases the probability of a *further* such occurrence. It is easy to see that only the numbers of reds and of blacks previously drawn, and not their sequence, determine the probabilities for the next drawing.

In the Polya Urn Scheme, success is rewarded by an increased chance of further success, but failure (*i.e.*, a black ball) is "punished" by an increased chance of further failure. Contagion, so to speak, is double-edged. In sociology, Merton calls this the Matthew Effect since "unto him that hath is given and unto him that hath not is taken away, even that which he hath." In fact, as we shall show, many of the known empirical data can be made to agree with a law which proceeds from the first part of the verse, but without the negative feedback of the second part. A trivial modification can make the effect of contagion single-edged so that success *increases* the chance of further success, but failure has no subsequent effect in changing probabilities. In effect, for many of the applications discussed, failure does not constitute an event as does success. Rather it must be accorded the status of a "non-event"; thus lack of publication is a non-event and only publication becomes a markable event. It seems to me that this difference between single- and double-edged contagious after-effects is the criterion on which one may decide whether the negative binomial or Cumulative Advantage distributions should apply.

For this urn model we shall suppose that after each drawing the ball is replaced; if a red is drawn then c red balls are added, but if a black is drawn no extra balls are put in the urn. If we start with b black balls and r red, the conditional probability of success after n previous successes will be $(r+nc)/(b+r+nc)$ and the corresponding conditional probability of failure will be $b/(b+r+nc)$.

The simplest case of this Cumulative Advantage Urn Scheme is to take $b = r = c = 1$ for, apart from the multiplicative constant c, any other starting configuration can be regarded as the state after some number of successes have occurred either under this rule, or in the complementary game with red and black reversed. The conditional chance of success after n successes is $(1+n)/(2+n)$ and the conditional chance of failure is $1/(2+n)$. Consider now, again as the most simple case, a population of N such urns, each originally in its ground state of $n = 0$. Going through the urns one by one we make drawings, continuing with the *same* urn as long as it produces successes and terminating and going on to *another* urn at the first failure. By the time all the N urns have been played, there will result a simple distribution which may be taken as a limiting case of the cumulative advantage process.

Half the urns will have produced a failure on their first drawing and therefore remain in that state. Of the half population that produce a first success, two-thirds will be successful again and one-third will fail and terminate; of the $N/3$ with two successes, three-quarters will succeed yet again, and so on. The expected number of urns with at *least* n successes by the end of the game will be $N/(n+1)$ and the expected number with *exactly* n successes is

$$\frac{N}{n+1} - \frac{N}{n+2} = \frac{N}{(n+1)(n+2)} .$$

Introducing a standard notation for the number of members of the population with exactly n successes, $N(n)$, and for the cumulative number of those with n or more successes, $S(n)$, we have:

$$N(n) = \frac{N}{(n+1)(n+2)} \tag{1}$$

$$S(n) = \frac{N}{n+1} \tag{2}$$

Setting $S(n) = 1$, it is seen that there is just one urn, that with the highest score of successes, in the range $N < n+1 < \infty$, thus the highest possible score is at least $N - 1$. Proceeding similiarly by setting $S(n) = m$, we may show that the mth highest score is in the range

$$N/m < n+1 < N/(m-1).$$

The total number of successes over the whole population may be found by summing $S(n)$ from the highest scorer downwards, though we must be careful to exclude the infinity that would be introduced by the upper limit of the highest scoring member of the population. With this exclusion, the sum of the lower limits is a harmonic series

$$\sum_{n=1}^{n=N-1} S(n) = N \sum_{n=2}^{n=N} \frac{1}{n} = N(C - 1 + \log_e N + o\,[\tfrac{1}{n}]),$$

where C is Euler's Constant, 0.577215665. . . The mean minimum score of successes per urn increases logarithmically as the size of the population. The mean value for the maxima, excluding only the highest scorer, is a similar expression, substituting $N+1$ for N. It is, therefore, convenient for most purposes to take this value, which has been derived as representing a mean value, the indeterminate upper bound of the highest score being excluded.

• The Cumulative Advantage Distribution

The Urn Scheme which has been discussed is, as has been stated, the simplest case of a Cumulative Advantage principle. A more general case, corresponding to a situation in which the population does not commence in a uniform ground state, may be studied better by considering a stochastic pure birth process, the general case which has been given by Feller [(*2*) I:448]. For this we shall suppose the population to consist of a number of individuals, each of whom is in a state that can be characterized by a single number, n, the total of "successes" thus far achieved. We further suppose that transitions or jumps can occur only by the incidence of a further success which transforms the individual from a state n to a state $n+1$, but never in the reverse direction.

A picturesque illustration of such a stochastic process would be a sort of "negative" radioactivity in which atoms of atomic (success) weight n could decay (negatively!) into atoms of the next higher weight. At any time the population consists of atoms distributed over the entire spectrum of atomic weights, all decaying upwards in the series at appropriate rates. If no new atoms at ground state were added, the entire population would decay towards infinite atomic weight; but if new ground state atoms are *continually* added, it can be shown that a stable population distribution can be produced, provided that a principle of cumulative advantage governs the transitions. The simplest expression of such a principle is to suppose that successes fall equally on the heads of all *previous* successes, so the frequency of transitions from state n to $n+1$ will be proportional to n.

This is fine if one begins from a ground state of unity rather than zero, as one does for example in considering a population of authors of n papers, an "author" of zero papers being undefined. In such a case, however, of the citedness of papers, it is useful to attach some meaning to the transition from zero citations to the first. If we wish to retain the proportionality between transition frequency and the state number, the number of successes must be counted as one *more* than the number of citations. In a way, this is tantamount of considering the *original* publication as the first citation, which seems reasonable, though one may wish the option of counting publication as some other number of successes, more or less than unity. Such a modification is relatively easy to make after the theory is developed for one may suppose the actual number of citations to be $n+k$, where n is the number of successes, and k is an arbitrary constant, presumably in the region of unity. Even more generally, we may set $n = ax+b$ where x is some other arbitrary state descriptor such as income in dollars.

Consider then a population of P individuals, of which a fraction $f(n)$ are in state n, where

$$\sum_{1}^{\infty} f(n) = 1,$$

and the mean number of previous successes

$$\sum_{1}^{\infty} nf(n) = R.$$

Now suppose that a small number dP of new individuals are added to the population, and with them RdP new successes are sprinkled at random over all members. Since these new successes are to be sprinkled evenly over the RP previous successes, there are dP/P new successes per previous one, and for the class of $Pf(n)$ individuals with n previous successes each, there will, therefore, be $nPf(n).dP/P$ new successes, and therefore transitions from this nth state to the $(n+1)$th. There must be therefore $nf(n)dP$ transitions out of the nth state, and there will be also $(n-1)f(n-1)dP$ transitions into it, from the class below receiving its quota of new successes. The

change in the number of individuals in the nth state is therefore

$$\frac{d}{dP}Pf(n) = -nf(n) + (n-1)f(n-1) \qquad n > 1$$
$$= -f(1) + 1 \qquad n = 1 \quad (3)$$

so that $P\frac{d}{dP}f(n) = -(n+1)f(n) + (n-1)f(n-1) \quad n > 1$
$$= -2f(1) + 1 \qquad n = 1 \quad (4)$$

and the distribution over the states is defined by this series of difference-differential equations. It can readily be seen that for a stable distribution, for which $f(n)$ is independent of P, the left hand side of Equation (4) becomes zero, and one may solve recursively

$$f(n) = \frac{n-1}{n+1}f(n-1) = \frac{n-1}{n+1}\cdot\frac{n-2}{n}\cdot\frac{n-3}{n-1}\cdots\frac{1}{3}\cdot\frac{1}{2} = \frac{1}{n(n+1)} \quad (5)$$

which is the form found for the Urn Model.

Suppose now that the distribution is slowly changing with P, and take it to be separable as the product of a function of P independent of n, and a function of n independent of P, thus

$$f(n) = F(P)g(n). \quad (6)$$

Substituting this in Equation (4) we get

$$\frac{P}{F(P)}\frac{dF(P)}{dP} = \frac{(n+1)g(n) + (n-1)g(n-1)}{g(n)} \qquad n > 1$$
$$= \frac{-2g(1)+1}{g(1)} \qquad n = 1 \quad (7)$$

Since the variables have been separated, both sides of Equation (7) must be constant of all P and for all n; let us call this constant m. It then follows that

$$\frac{P}{F(P)}\frac{dF(P)}{dP} = m \quad (8)$$

therefore $F(P) = CP^m$

and $$\frac{-(n+1)g(n) + (n-1)g(n-1)}{g(n)} = m \qquad n > 1$$
$$\frac{-2g(1)+1}{g(1)} = m \qquad n = 1 \quad (9)$$

Therefore

$$g(n) = \frac{n-1}{n+1+m}g(n-1)$$
$$= \frac{n-1}{n+1+m}\cdot\frac{n-2}{n+m}\cdots\frac{1}{3+m}\cdot\frac{1}{2+m}$$
$$= \frac{(n-1)!(m+1)!}{(n+1+m)!} \quad (10)$$

This last result may be expressed most elegantly by using the notation of the Beta Function, otherwise known as Euler's First Integral,

$$B(a,b) = B(b,a) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx =$$
$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \frac{(a-1)!(b-1)!}{(a+b-1)!}$$

So that from Equations (6), (8) and (10) we have a solution of Equation (3) as

$$f(n) = CP^m B(n,m+2) \quad (11)^*$$

and hence for any particular value of P, the distribution we have considered is proportional to the values of the Beta Function $B(n,m+2)$. This may be expressed in the usual way as a discrete probability density by chosing the constant of proportionality so that

$$\sum_1^\infty f(n) = 1,$$

and on this basis we propose as the Cumulative Advantage Distribution, the density

$$f^*(n) = (m+1)\,B(n, m+2)$$

though in most cases it will prove more convenient to use the unnormalized form and work directly with tables of the Beta Function.

The Beta Function, of course, assumes simple forms for integral values of a and b. It should be noted for example that

$B(n,1) = 1/n$
$B(n,2) = 1/n(n+1)$
$B(n,3) = 2/n(n+1)(n+2)$
$B(n,4) = 6/n(n+1)(n+2)(n+3)$, etc.

*A linear combination of such solutions for any values of m constitutes a more general solution, but I see at present no physical significance in such a compound expression.

We tabulate the function for integral values of the arguments by taking $B(1,b) = 1/b$ and then calculating successively

$B(2,b) = B(1,b)/(b+1), B(3,b) = B(2,b) \cdot 2/(b+2), B(4,b) = B(3,b) \cdot 3/(b+3)$, etc.

• Properties of the Beta Function

It should be noted that $f(n)$ contains only the single parameter m, the simple limiting case discussed above in the Cumulative Advantage Urn Model corresponding to $m = 0$. The distribution is therefore remarkably simple in form, having one free parameter like the Poisson distribution, rather than two or more as in other contagious probability forms like the negative binomial. The properties of the Beta Function are, of course, well known, though their use in this context seems not to be widespread since it appears that no previous extensive tabulation of the function has been published.

Perhaps the most important property for our purposes is that

$$B(a,b) - B(a+1,b) = B(a,b+1) \tag{12}$$

so that the values of $B(a,b+1)$ are the first differences of $B(a,b)$ and, in general, $B(a,b+c)$ for the run of values of a are the cth differences of $B(a,b)$. A table of the Beta Function for integral values of a and b has therefore for $b = 1$, a table of the reciprocals of the nature numbers, and for higher values of b, the bth differences of this harmonic series. Conversely, since differences are taken by increasing the value of b, we get a sum by decreasing it. Each run of $B(a,b)$ gives the sum from infinity to a of $B(a,b+1)$. Since the Cumulative Advantage Distribution is given in Equation (11) by the Beta Function with $b = m+2$ for the relative number of the population with *exactly* $a = n$ successes, the cumulative sum for all $a = n$ must be given by the adjacent tabulation for $b = m+1$. In a tabulation for integral m (see Table 1) we shall have therefore, in adjacent columns, the frequency distribution on the right and the cumulative frequency distribution on the left.

As a special case it may be noted that the top number in any column, *i.e.*, the value $B(1,b)$, is the sum of the entire column to the right of it, *i.e.*,

$$\sum_{a} B(a,b+1).$$

Because, as was noted above in the Urn Model, the total number of successes is the sum of the *cumulative* frequencies, the top entry to the left of the pair of columns just mentioned gives this sum of successes, all numbers being, of course, scaled to the total population. In all then, for any particular value of $m > 1$, we shall be concerned with three adjacent columns of the Beta Function table:

$B(1,m)$ gives the total number of successes,

$B(1,m+1)$ gives the size of the population (except for a factor if we wish to make this equal to P), and subsequent entries in this column $B(n,m+1)$ given the cumulative frequency, *i.e.*, the number of members with success scores of at least n.

$B(n,m+2)$ gives the actual frequencies, *i.e.*, the number of members with success scores exactly m.

Since the size of the population is $B(1,m+1) = 1/(m+1)$, we may norm to a population P by multiplying all entries by $P(m+1)$. As leading results from such a normalization, we note that the total number of successes becomes $P(m+1)B(1,m = P(m+1)/m$, and the mean is therefore $1+1/m$; and that the proportion of individuals with exactly one success is $(m+1)B(1,m+2) = (m+1)/(m+2)$. These results may be used as estimators for the parameter m, but for this another very useful indication is that for large values of a,

$$B(a,b) = \frac{(b-1)!}{(a+b-1)(a+b-2)\cdots(a+1)a} \simeq \frac{(b-1)!}{(a+\frac{b-1}{2})^b} \tag{13}$$

So, rather more approximately,

$$B(a,b) = (b-1)!a^{-b} \tag{14}$$

and as a result, for reasonably large values of n, the Cumulative Advantage Distribution follows an inverse power law with exponent $m+2$; and the cumulative distribution is also an inverse power law, but with exponent $m+1$. This is indeed the characteristic of several very well known empirical laws of social science, including bibliometrics (for which see the splendid critical summary by Fairthorne (*14*) and the recent literature review by Narin (*15*). In most cases such as the Pareto and Lotka distributions, it appears that m is small so that one is not far distant from the limiting case $m = 0$ which was discussed in the Urn Model.

• The Limiting Case—Lotka and Bradford

It must be observed that in this limiting case we can no longer use an estimation from the mean, since the cumulative distribution $B(n,1)$ is the harmonic series $1/n$ and its sum is divergent. In this case, we must argue that

Table 1

a=	b= 1	b= 2	b= 3	b= 4	b= 5	b= 6	b= 7	b= 8	b= 9	b= 10	b= 11
1	10000000	5000000	3333333	2500000	2000000	1666667	1428571	1250000	1111111	1000000	909091
2	5000000	1666667	833333	500000	333333	238095	178571	138889	111111	90909	75758
3	3333333	833333	333333	166667	95238	59524	39683	27778	20202	15152	11655
4	2500000	500000	166667	71429	35714	19841	11905	7576	5051	3497	2498
5	2000000	333333	95238	35714	15873	7937	4329	2525	1554	999	666
6	1666667	238095	58524	19841	7937	3608	1804	971	555	333	208
7	1428571	178571	39683	11905	4329	1804	833	416	222	125	73
8	1250000	138889	27778	7576	2525	971	416	194	97	51	29
9	1111111	111111	20202	5051	1554	555	222	97	46	23	12
10	1000000	90909	15152	3497	999	333	125	51	23	11	5
11	909091	75758	11655	2498	666	208	73	29	12	5	3
12	833333	64103	9158	1832	458	135	45	17	7	3	1
13	769231	54945	7326	1374	323	90	28	10	4	2	1
14	714286	47619	5952	1050	233	61	18	6	2	1	
15	666667	41667	4902	817	172	43	12	4	1	1	
16	625000	36765	4085	645	129	31	8	3	1		
17	588235	32680	3440	516	98	22	6	2	1		
18	555556	29240	2924	418	76	17	4	1			
19	526316	26316	2506	342	59	12	3	1			
20	500000	23810	2165	282	47	9	2	1			
21	476190	21645	1882	235	38	7	2				
22	454545	19763	1647	198	30	6	1				
23	434783	18116	1449	167	25	4	1				
24	416667	16667	1282	142	20	4	1				
25	400000	15385	1140	122	17	3	1				
26	384615	14245	1018	105	14	2					
27	370370	13228	912	91	12	2					
28	357143	12315	821	79	10	2					
29	344828	11494	742	70	8	1					
30	333333	10753	672	61	7	1					
31	322581	10081	611	54	6	1					
32	312500	9470	557	48	5	1					
33	303030	8913	509	42	5	1					
34	294118	8403	467	38	4	1					
35	285714	7937	429	34	3						
36	277778	7508	395	30	3						
37	270270	7112	365	27	3						
38	263158	6748	337	25	2						
39	256410	6410	313	22	2						
40	250000	6098	290	20	2						
41	243902	5807	270	18	2						
42	238095	5537	252	17	1						
43	232558	5285	235	15	1						
44	227273	5051	220	14	1						
45	222222	4831	206	13	1						
46	217391	4625	193	12	1						
47	212766	4433	181	11	1						
48	208333	4252	170	10	1						
49	204082	4082	160	9	1						
50	200000	3922	151	9	1						
51	196078	3771	142	8	1						
52	192308	3628	134	7	1						
53	188679	3494	127	7							
54	185185	3367	120	6							
55	181818	3247	114	6							
56	178571	3133	108	5							
57	175439	3025	103	5							
58	172414	2922	97	5							
59	169492	2825	93	4							
60	166667	2732	88	4							
61	163934	2644	84	4							
62	161290	2560	80	4							
63	158730	2480	76	3							
64	156250	2404	73	3							
65	153846	2331	70	3							
66	151515	2261	67	3							
67	149254	2195	64	3							
68	147059	2131									

Table 2

a=	b= 1.0	b= 1.1	b= 1.2	b= 1.3	b= 1.4	b= 1.5	b= 1.6	b= 1.7	b= 1.8	b= 1.9	b= 2.0	b= 2.1	b= 2.2	b= 2.3
1	10000000	9090909	8333333	7692307	7142857	6666666	6249999	5882352	5555555	5263157	5000000	4761904	4545454	4347826
2	5000000	4329004	3787879	3344481	2976190	2666666	2403846	2178649	1984127	1814882	1666667	1536098	1420454	1317523
3	3333333	2792906	2367424	2026958	1750700	1523809	1335470	1177648	1044277	930709	833333	749316	676407	612801
4	2500000	2043590	1691017	1414157	1193659	1015873	870959	751690	652673	569822	500000	440774	390235	346869
5	2000000	1602815	1300782	1067288	884192	738817	622113	527502	450120	386320	333333	289032	251764	220234
6	1666667	1313783	1049018	847054	690775	568321	471298	393658	330970	279942	238095	203544	174836	150845
7	1428571	1110239	874182	696209	560088	454656	372077	306747	254592	212614	178571	150773	127929	109045
8	1250000	959466	746253	587164	466740	374423	302854	246808	202517	167225	138889	115979	97337	82077
9	1111111	843487	648915	505087	397225	315304	252378	203553	165320	135131	111111	91864	76343	63749
10	1000000	751622	572572	441339	343753	270260	214283	171213	137766	111576	90909	74485	61347	50773
11	909091	677137	511225	390565	301538	235009	184727	146335	116751	93761	75758	61558	50284	41279
12	833333	615579	460941	349286	267493	206808	161270	126747	100333	79952	64103	51690	41904	34141
13	769231	563889	419037	315145	239546	183829	142297	111020	87246	69023	54945	43991	35412	28650
14	714286	519898	383626	286496	216257	164812	126702	98181	76635	60221	47619	37873	30286	24343
15	666667	482024	353339	262153	196597	148863	113707	87550	67905	53025	41667	32933	26173	20908
16	625000	449091	327166	241245	179814	135330	102748	78637	60629	47064	36765	28889	22826	18128
17	588235	420202	304341	223117	165347	123730	93407	71085	54498	42068	32680	25537	20066	15850
18	555556	394665	284274	207267	152766	113698	85372	64622	49280	37839	29240	22729	17767	13961
19	526316	371936	266507	193306	141741	104952	78403	59046	44800	34226	26316	20355	15832	12379
20	500000	351581	250675	180927	132014	97273	72313	54197	40923	31115	23810	18329	14189	11042
21	476190	333252	236486	169884	123378	90486	66957	49951	37544	28415	21615	16587	12783	9904
22	454545	316665	223703	159981	115667	84454	62216	46210	34580	26058	19763	15079	11571	8926
23	434783	301586	212132	151055	108746	79063	57998	42895	31965	23986	18116	13765	10519	8081
24	416667	287820	201613	142974	102507	74222	54226	39943	29645	22156	16667	12614	9601	7346
25	400000	275207	192012	135627	96857	69856	50837	37301	27577	20531	15385	11599	8794	6704
26	384615	263608	182218	128923	91720	65902	47779	34926	25724	19080	14245	10700	8083	6139
27	340370	252908	175135	122784	87034	62308	45009	32782	24059	17781	13228	9900	7453	5640
28	357143	243008	167682	117144	82744	59028	42491	30841	22555	16612	12315	9186	6891	5198
29	344828	233822	160791	111946	78803	56027	40194	29075	21193	15556	11494	8545	6389	4803
30	333333	225277	154402	107143	75174	53271	38093	27465	19954	14600	10753	7968	5939	4450
31	322581	217309	148464	102693	71823	50735	36164	15992	18825	13730	10081	7447	5533	4133
32	312500	209862	142931	96857	68719	48393	34389	24641	17792	12937	9470	6974	5166	3848
33	303030	202888	137765	84712	65839	46226	32752	23398	16844	12212	8913	6545	4834	3590
34	294118	196343	132931	91123	63159	44216	31237	22252	15973	11547	8403	6153	4532	3356
35	285714	190190	128399	87767	60661	42348	29833	21192	15170	10936	7937	5795	4256	3143
36	277778	184395	124143	84624	58328	40608	28529	20211	14428	10373	7508	5467	4005	2949
37	270270	178927	120138	81674	56145	38983	27315	19299	13741	9853	7112	5166	3774	2772
38	263158	173761	116364	78902	54098	37465	26183	18451	13103	9372	6748	4888	3562	2610
39	256410	168873	112802	76292	52176	36042	25125	17661	12511	8926	6410	4632	3367	2461
40	250000	164241	109435	73831	50368	34707	24135	16924	11959	8511	6098	4396	3187	2324
41	243902	159845	106247	71507	48664	33453	23206	16234	11444	8125	5807	4176	3021	2198
42	238095	155668	103226	69309	47058	32272	22335	15587	10963	7765	5537	3973	2867	2081
43	232558	151695	100359	67229	45540	31159	21515	14981	10512	7429	5285	3784	2725	1973
44	227273	147912	97634	65256	44104	30109	20743	14411	10090	7115	5051	3608	2592	1873
45	222222	144304	95042	63383	42744	29116	20016	13875	9693	6820	4831	3443	2469	1780
46	217391	140861	92573	61603	41454	28177	19328	13370	9320	6544	4625	3290	2354	1693
47	212766	137571	90220	59910	40230	27287	18679	12894	8969	6284	4433	3146	2246	1612

some maximum value of n must exist, as before that for which the cumulative number of cases becomes unity; this leads again to a mean given by $C + \log_e P$. If in this case we take a running cumulative sum of the score of successes, starting at the maximum, and consider it as a function of rank, one gets the same sort of logarithmic relationship. From Equation (2), again with a proviso to exclude the infinity due to the upper range of the highest score, we count the highest score as P-1 and the rth highest as $P/r - 1$. The sum of the scores from the top to the rth individual is therefore:

$$Q(r) = (P\text{-}1)+(P/2\text{-}1)+(P/3\text{-}1)\ldots(P/r\text{-}1)$$
$$= P(1+1/2+1/3\ldots 1/r)\text{-}r \quad (15)$$
$$= P(C+\log_e r)\text{-}r. \quad (16)$$

If r is large compared with e^C=1.781 and small compared with P, this last equation becomes

$$Q(r) \simeq P\log_e r \quad (17)$$

so that the total success score is proportional to the logarithm of the rank from the top. This is well known as the basis from which one may derive the Bradford Distribution (*15, 16, 17*) in which geometric "zone" increases in rank to correspond to arithemetic increments in the total score. In the simplest two-zone form, it states that half the successes are due to the highest scoring elite comprising $\sqrt{P}$ of the individuals. Though we did not realize it at the time of publication (*18*), this turns out to be precisely the same mathematical basis for the so-called Price Law that asserts, on the basis of the Lotka distribution, that the top $\sqrt{P}$ authors will produce at least half the total papers published by the population P.

We see, therefore, that two of the most popular empirical laws of bibliometrics and some other findings can be derived immediately from the underlying theory of Cumulative Advantage in its limiting case $m = 0$, and their forms have much more general application than had been supposed. Insofar as the empirical laws do not quite fit all of the data, a very probable cause is that, in such cases, the limit is approached but not reached, so that m is small but non-zero. Two additional small modifications for practical use may be noted: for the usual step form of the Bradford graph, it is better to use the actual harmonic series terms of Equation (16) rather than the logarithmic approximation of Equation (16); and the Groos (*19*) "Droop", having no underlying theoretical basis, is simply due to the poor ability of gathering individuals with very low scores. It may, in fact, be explained as a Poisson probability for rare events (successes) that would lead one to gather only a fraction $1-e^{-p}$ of individuals when there is only a small expectation p, that is, only a few successes each. For small values of n one should therefore apply this correction.

All such modifications notwithstanding, it follows from Equation (17) that if we have a population in which there operates—for whatever cause—a Cumulative Advantage Process in its limiting case;

$$Q(r)/Q(P) = \log r/\log P. \quad (18)$$

Thus for example, if we suppose, as turns out from the other data to be reasonable, that citation of a journal is governed by such a process, we may calculate the fraction of all citations to journals or papers that will be collected by starting from the r most cited journals instead of the total population of P journals. Barr (*20*) estimated that there were about 26,000 journals of interest to scientists and technologists in 1966; for that same year the *Science Citation Index*® used 1573 source journals. If we suppose that these were all the most highly ranked journals, they would yield log 1573/log 26000 = 0.72 of all citable papers of that year, which may be regarded as the built-in efficiency of an operation based on a selection rather than the entire population.

In the same way, it is trivial mathematically, but by no means so economically, to observe that if we have, say one million individuals of the highest rank in a Cumulative Advantage population in its limiting case, the doubling of the population by the addition of the next highest ranked million will increase the total bulk of successes by only log 2/log $10^6 \simeq 5\%$. Thus, if "successes" be book use in a library, citations to a body of literature, frequencies of word use or incomes of the rich, once one has a large body of individuals selected by their being of the highest ranks in success rate, the pay-off for *adding* to the population becomes rather small. Since it happens that the cost of managing or holding a collection is proportional to more than the first power of the size of the collection, the acquisition of an additional 5 percent of value in the return is paid for by more than a doubling in cost. We, therefore, see the operation of a very powerful principle of marginal utilities in such cumulative advantage situations. It is the marginal economy, of course, that produces the force leading one to the elitism (*21*) of small libraries, vocabularies and document collections and small social groups of those with many successes behind them. For the *Science Citation Index*®, the operation would not be possible if it were not for the Cumulative Advantage situation. Taking as sources the most obvious and best 6 percent or so of all citable journals one can, as has been shown, acquire citations to about 72 percent of all citable papers.

• Application of Cumulative Advantage Theory to Lotka's Law

Although much comment has been made about Lotka's Law of Scientific Productivity, including interesting theoretical pseudo-derivations by Shockley (*22*) and by Zener (*23*), its very simplicity seems to have militated against the collection of empirical data that would give a parameter m in terms of the mean value or by some other way. Lotka's (*24*) original data, tested only for relatively small n, was based partly on lifetime productivities and partly on the ten year indexes. In a way, it is rather misleading, for one is led to suppose that the $1/n^2$ law can be identified with the limiting case of the Cumulative Advantage Distribution for $m = 0$, for which $N(n) = 1/n(n+1)$. It is possible if scientists did not have a finite research lifetime, and if we had a total bibliography over all time, such a law might hold. For Lotka's test, however, it is only necessary that the first few values for small n fit the inverse square law so that

$$N(1)/N(2) = 2^2,\ N(1)/N(3) = 3^2, \text{ etc.}$$

More crucially, as was shown by Price (*25*) p. 46-48, the ranking of the most productive authors, in Lotka's data and in many other examples, follows the law $S(n) = c/n^2$ which leads one to suppose that m is in the region of unity rather than zero.

If the value of m were exactly unity (which it doubtless is not), we would have

$$N(n) = B(n,3) = 2/n(n+1)(n+2),\ S(n) = B(n,2) = 1/n(n+1),$$

$S(1) = 1/2$ and $Q(1) = B(1,1) = 1$, so that the mean

$Q(1)/S(1) = 2$. It follows therefore that $N(1)/N(2) = 4$ exactly as required, and $N(1)/N(3) = 10$ which is close to 3^2. A still more crucial test for the distribution in the tail is given by the computation of the highest scoring member by setting $PS(n) = 1$. For $m = 1$, taking a population of the order of a million authors we must have $n(n\text{-}1) \simeq n^2 = 10^6$ which gives a maximum score of $n = 1000$, in reasonable accordance with the order of magnitude of existing world records. Another reasonable estimate is to be had from the proportion of single paper authors, $N(1)/S(1) = 2/3$ for $m = 1$, which is somewhat higher than the approximately 60 found empirically and theoretically by Lotka.

The small discrepancy could be reduced by using a slightly smaller value for m, say 0.7, rather than unity. This value makes the maximum score for a million authors $(10^6)^{1/1.7} = 3400$ in close agreement with the remarkable actual world record of 3904 papers attributed to T.D.A. Cockerell (1866-1948), an entomologist of the University of Colorado. Also for this value we have $N(1)/N(2) = 3.7$ and $N(1)/N(3)$ 8.7, both rather

	$m = 1.0$		$m = 0.7$	
n	$N(n)$	$S(n)$	$N(n)$	$S(n)$
1	.333	.500	.370	.588
2	.083	.166	.100	.218
3	.033	.083	.043	.118

near to the values of 4 and 9 required by the inverse square law, and the mean value of productivity becomes

$$1+1/0.7 = 2.43.$$

The actual value of m is likely to vary with the sampling used in any particular case. For modern data there must be much uncertainty because we have as yet no adequate model or theory for the attribution of credit in the case of multi-author collaborative papers; there remains also some doubt as to the effect of finite lifetime on author productivity. Such difficulties notwithstanding, it is clear from the only extensive demographic study of authorship thus far, that by Price and Gürsey (*26*), that authors differ not so much in their *rate* of publication of papers, but in the span of *time* they spend at the publication front. The greater number of those arriving at the research front emit only their initial publications and are therefore to be considered as "transients"; they number perhaps two-thirds of all authors. This is just what would be expected in a Cumulative Advantage process with m near unity; it corresponds in fact to an Urn Model where one begins with one red ball of success and two blacks for failure. Two-thirds of all first choices will result in failure and the end of that particular "game" of publication, but for the one-third that *succeed* there will be an even chance of a *second* success, and thereafter the chance will approach unity asymptotically. It seems therefore that in all qualitative aspects and also quantitatively the Cumulative Advantage Distribution and process account for journal distributions and author productivities.

• Application to Citation Data

By far the most stringent test of the Cumulative Advantage principle is the application to citation data where we have a wealth of computer-generated counts instead of the short series of hand counts for productivity. Though the data are rich, the empirical generalizations and underlying theoretical constructs have been sparse. Price (*27*) reported that the number of papers cited n times in a year followed an inverse power law (a

Zipf Law) with exponent in the range 2.5-3.0, and I. Yermish (*28*) has since shown that an exponent of 3.036 gives a correlation of 0.9937 with the data for 1972. The only far-reaching theoretical analysis known to me is by Charles J. Crowley (*29*) where it is supposed that a negative binomial distribution would have the required properties, and this is unfortunately tested on very limited data. One other previous contribution deserves special mention even though its ingenious and elegant mathematics were introduced in connection with the Bradford Law rather than citation data. Brookes (*30*) showed that the sampling of a truncated hyperbolic distribution could be accounted for by an analogue of the Taylor's Theorem, a result which I conjecture may be equivalent to convoluting or changing the value of m to $m+h$ in the Beta Function, $B(n,2+m)$ of the Cumulative Advantage Distribution.

In all that follows, it must be remembered that for citations there is a problem associated with the zero ground state. We have supposed above that the *publication* of a paper might be counted as its first citation "success." If it is *not*, then the value of n in the Beta Function must be taken as the number of citation successes plus k, where k is some parameter different from unity. This will affect the values of $N(n)$ and $S(n)$ for small values of n only, and it may be used as an additional parameter to adjust such values to agree with the empirical data. Because of the possibility of such a 'fudge' factor being used, we shall not rest any vital tests upon such values for small n, and provisionally k is set at unity.

For the total citation network for all time, it is clear that the maximum number of successes per paper is equal to the total population of papers, for it has its own publication and can be cited by all other papers. It follows then that the average number of successes, which must be one more than the average number of references from each paper to the same journal literature, must be $C\text{-}\log_e P$ which is here tabulated for typical values of population:

P	*Mean Successes*	*Refs/Paper*
1,000	7.48	6.48
10,000	9.79	8.79
100,000	12.09	11.09
1,000,000	14.39	13.39
10,000,000	16.70	15.70
100,000,000	19.00	18.00

Since there are, in fact, about 13 references per article today across the whole field of literature, this provides an over-estimate, especially since many of these are to non-journal items outside the source network. What has probably happened also is an overestimation of the maximum possible success rate; even the most cited paper, that of Lowry 1951, is cited only by about 1 percent of all papers. With this correction, a rate near the actual ten journal references per article gives a total archive of 3,400,000 papers which is perhaps a slight *under*estimate. It may well be that this tells us that the scientific literature consists not of a unity, but of some hundreds of nearly autonomous subfields. Alternatively it might be due to some process whereby only a fraction, *ca.* 10/13 of all successes, are actually manifested as citations.

A remarkable consequence of this derivation is that each doubling of the archive should add $\log_e 2 = 0.7$ refs/ paper. Since the total population of papers grows exponentially with a doubling time of *ca.* ten years, we should expect that the number of refs/paper must have been increasing by unity about every 15 years, a fact that agrees well with the much smaller incidence of citation in the early years of this century. I do not believe this postdiction of the bibliometric record has ever before been stated or observed.

For an actual citation index, such as those published on a quarterly, annual and quinquennial basis by the Institute for Scientific Information, we must take into account that this is a sampling in two distinct ways. First, we have the fact that it is based upon a selection of the largest and most cited journals, and, as we have seen, if this in itself is governed by a Cumulative Advantage Distribution, it is probable that this collects a fraction of about 72 percent of the entire cited literature. Secondly, the index is based upon only a restricted range over time of the previous literature, which we shall suppose to be in exponential growth at a rate K, which is approximately 0.07 per year. Letting the number of journal references per source article be R, it follows that in a one-year citation index there will be KR references back per article existing in the corpus. As has long been known [see Price (*27*)], this works out to about one citation per article per year. Since in the Cumulative Advantage Distribution we have a mean of $1+1/m$ successes/item, it follows that the mean number of citations M is one less than this, $1/m$. We are led therefore to suppose that a one year citation index must have m near unity, and it will therefore follow from Equation (14), for large n, an inverse cube law, $1/n^3$. For a many-year index the mean number of citations per corpus paper will be larger, so that m is small, and one gets closer to the inverse square law which is the limiting case. For a correspondingly shorter period, the value of m must be larger than unity; for example, it must be that $m \simeq 4$ for a quarterly index, so that the distribution for large n will fall off as $1/n^6$. It follows also that the cumulative distributions will fall off as $1/n^2$ for the annual, nearly $1/n$ (perhaps $1/n^{1.2}$) for the quinquennial, and $1/n^5$ for the quarterly citation indexes.

Table 3. Empirical data for number of papers cited n times in *Science Citation Index*®

n	Quarterly	Σ	Annual	Σ	6-year Cum	Σ
		10,000		10,000		10,000
1	7,968.2		7,280.9		6,689.7	
		2,031.8		2,719.1		3.310.3
2	1,223.0		1,328.9		1,260.0	
		808.8		1,390.2		2,050.3
3	532.7		550.5		555.2	
		276.1		839.7		1,495.1
4	154.3		312.3		329.9	
		121.8		527.4		1,165.2
5	66.6		154.9		225.3	
		55.2		372.5		939.2
6	24.4		97.5		181.8	
		30.8		275.0		758.1
7	6.5		64.3		114.3	
		24.3		210.7		643.8
8	8.1		45.2		90.1	
		16.2		165.5		553.7
9	3.2		31.4		88.5	
		13.0		134.1		465.2
10	1.6		24.4		69.2	
		11.4		109.7		396.0
11	0		18.3		56.3	
		11.4		91.4		339.7
12	1.6		14.2		35.4	
		9.8		77.2		304.3
13	1.6		11.0		24.1	
		8.2		66.2		280.2
14	3.2		8.9		40.2	
		5.0		57.3		240.0
15	0		7.3		24.1	
		5.0		50.0		215.9
16	0		6.1		20.9	
		5.0		43.9		195.0
17	1.6		5.2		9.7	
		3.4		38.7		185.3
18	0		4.2		22.5	
		3.4		34.5		162.8
19	0		3.8		9.7	
		3.4		30.7		153.1
20	0		3.0		24.1	
		3.4		27.7		129.0
>20	3.4		27.7		129.0	
Average cites/paper cited	1.345		1.66		2.641	

Sources:
Quarterly, hand count of 6157 papers from 1975 index.
Annual, computer count of 1,882,864 papers from 1967 index.
5-Year Cum, hand count of 6214 papers from 1965-1969 index.

In an empirical test from machine-generated data for the annual and from hand-counts for the quarterly and quinquennial indexes (see Table 3 and Fig. 1), I find for the actual distributions exponents of 2.1 for five years, 3.2 for one year, and 5.3 for a quarter year, and independently for the cumulative distributions, values of 1.4, 2.3 and 4.0 respectively. It would appear therefore that for an annual index we have $m \simeq 1.25$, and for the quinquennial it has an appropriate value about one-fifth of this, $m \simeq 0.25$, but for the quarterly where we should have $m \simeq 5.0$, we have in fact $m \simeq 3.15$ which may be due to the smallness of the sample counted, or it may have some deeper basis such as the constant k associated with the ground state transition not being unity.

Another interesting test of the citation data is afforded by the rank list of highly cited items. According to the tabulation by Garfield (*31*), the most cited paper, that of Lowry *et al* had 29,655 citations in the interval 1961-1972, followed by items with 6281, 5825, 5273 and 5054 citations in the same ten year period. The tenth most cited paper had 3621; the 20th, 2054; the 30th, 1695; the 40th, 1317 and the 50th, 1207. The trend follows approximately (except for the anomalous top item) a typical power law that the number of citations is $10{,}000/\text{rank}^{2/3}$, which is equivalent to a cumulative distribution proportional to $1/n^{1.5}$ suggesting that for a ten year index we should have $m \simeq 0.5$.

In this period the citation index contained a total of *ca.* 3,500,000 source items, and although the annual rate grew by about a factor of three, some of that was due to an expansion in the extent of the Institute for Scientific Information operation relative to the total available literature, so one may suppose that the actual corpus, growing at *ca.* 7 percent per annum, must have doubled in the period. The corpus at the end of the period therefore must have been about 7 million papers, so that Lowry's paper had been cited by more than four papers in every 1000, a surprisingly large fraction when one remembers that most scientific literature has nothing to do with organic chemistry. Quite apart from the trend of most cited items, we may use this highest score as an estimator, since from Equation (14) it follows that, setting the cumulative total of unity, the score of the highest cited item must be given by $7{,}000{,}000/n^{1+m} = 1$ and for $n = 29{,}655$ we have $m = 0.53$.

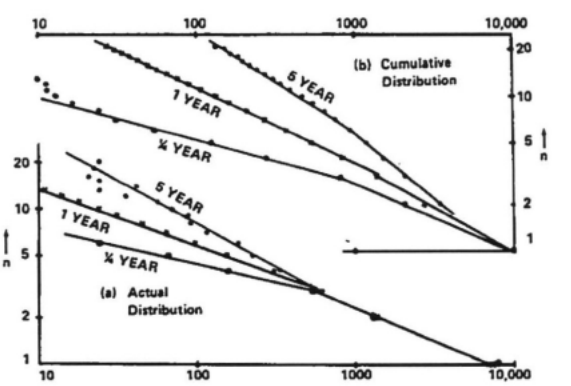

Fig. 1. Number of papers with (a) exactly and (b) at least n citations in ¼, 1, and 5-year indexes.

It should also be possible to estimate the value of m from the Garfield Constant (*32*), the number of citations per cited item. This parameter, G, has been slowly rising for the annual index from about 1.60 in the early 1960s to about 1.80 in the mid-1970s; for the 1965-1969 quinquennial index it was 2.55 and for quarterly indexes it is *ca.* 1.3. For the Cumulative Advantage Distribution the number of papers cited (*i.e.*, having at least two successes) is given by $S(2) = B(2,m+1) = 1/(m+1)(m+2)$ and the number of citations by

$$Q(2) = B(2,M) = 1/m(m+1),$$

hence $G = 1+2/m$ and the Garfield Constant is therefore one more than twice the mean number of citations per paper.

The fraction of all cited papers that occur in a citation index based upon a limited time interval T, compared with those in one based upon all time may be calculated by supposing that we have a Poisson process with an expectation RKT where R is the number of references/paper, and K is the growth rate. The fraction of papers not "hit" is e^{-RKT} so those "hit" by citation are $1-e^{-RKT}$ and combining this with the efficiency F (due to the journal selection) we have $M = F(1-e^{-RKT})$. For $F = 0.72$, $R = 9$, $K = 0.07$/year we get:

Span of Index in Years	Calculated Mean Citations = M	Calculated $G = 1+2M$
1/4	0.105	1.21
1	0.336	1.67
5	0.689	2.38

Alternatively, if we take G to be 2.55 for the quinquennial and 1.80 for the annual index, it follows from a simple numerical solution of the simultaneous equations that we must have $F = 0.79$ and $RK = 0.73$. It seems reasonable that the efficiency of the citation index should have improved by enlarging the journal list from 0.72 in 1966 to 0.79 in 1974 (indeed, computing from the current size of the roster of source journals log 2500/log 26,000 = 0.77), and for the value of RK we might have a growth rate 0.08 per year and 9.1 references to the journal literature per article. The agreement is surprisingly good though the values of M seem to correspond to $m+2$ rather than to m which I should expect.

• Citations as a Function of Time and Field

It should be noted that the Cumulative Advantage Distribution has been derived without explicit reference to time as a variable, and unlike such many-parameter functions as the negative binomial, it contains little scope for adjusting the constants to fit different fields or other sorts of population. The distribution depends only on the size of the population and the mean number of successes per item. We have supposed that successes generate future successes without any particular time scale. For an exponentially growing population, however, it is clear that if we start with RP successes there will be added KRP where K has a value of the order of 0.07 per year, and hence the exponential growth may be regarded as due to each success producing 0.07 new successes per year, or one further success every *ca.* 14 years. Such a concept is, of course, equivalent to a model explaining exponential growth as a tendency for successes to breed new successes at a constant birth rate.

The birth rate is however quite small. A paper needs to have some 14 previous successes before it reaches a rate of one new success per year. In view of this smallness of the constant, it is clear that the first few citations could not cumulate in such a slow fashion—it would take far too long for Lowry to become Lowry! What clearly happens is that immediately after publication, in as long as it takes for the work to become known (which may indeed precede publication through the use of informal communication) the paper is weighed by peers and in its incunable period produces a first pulse of citations which in most cases probably determines all future citation history; only rarely is there a much retarded discovery or subsequent rounds of rediscovery and new application.

In the absence of such anomalies the future citation history of a paper will depend only on the size of the initial pulse and since $dn = Kndt$, we have $n = n_0 e^{Kt}$ so that the number of citations is expected to grow at the same exponential rate as the literature, where t is the time elapsed from the initial pulse of n_0 citations. The pulse size is clearly a determinant of the effectiveness of the paper at the research front. It must thus be a measure of "quality," perhaps also containing a factor proportional to the size of the paper (in pages?).

An interesting point that must have further investigation is that there must be a feedback from the citations as a cumulative advantage process to the author's productivity which follows just the same pattern. We must suppose that the size of the initial pulses modulates the author's behavior and causes either a continuance of publication or a cessation. If the initial paper(s) are well received, the author's self-estimate, and also those who may determine things such as promotion and tenure, will enable the circumstances and motivation to be high enough to make another attempt. At subsequent determinations the cumulation of past success makes the threshold needed smaller so there arise larger and larger cumulating advantages leading to continuance in publica-

tion. Price and Gürsey (*33*) showed that transient authors have only a chance of about one in four of having their sole paper cited, non-core continuants have about 0.7 chance of being cited, and core continuants are almost all cited every year. The correlation between citation and productivity is therefore very high and most effective in the crucial decision to remain at the research front or retire from it.

The citation of papers, as has been shown, depends only on their nature and not on the number of papers available to cite them, though this too grows at the same exponential rate. This point was confirmed by Gomperts (*34*) who showed that in the field of vibrating plates the citations per paper were independent of the size of the citing literature, though of course the rate must vary from paper to paper. Looking back on the literature from any date however, by observing the references in source papers of that date, one must get a distribution that depends on the number of papers available for citation at any previous date. In 1965 I suggested that the time distribution of citations per cited paper should enable one to disentangle the literature growth from any obsolescence factor that might apply. Now, the cumulative advantage theory enables this to be done, for as has been seen the citations per cited paper (the Garfield Constant) is related to the mean M and therefore to the parameter m. We have $G = 1+2M = 1+2/m$. Taking the data from Price (*27*) we get the mean citations, and therefore for the amount the literature of that age is used relative to the use per paper of most recent literature, the following table:

Age in years	G	$1-G/\ 1-G_0$	Age in years	Relative use in percent
0	1.75	1.00	2	98
4	1.7	0.93	3	96
9	1.6	0.80	5	90
15	1.5	0.67	8.5	80
20	1.4	0.53	10	75
28	1.3	0.40	12	70
42	1.2	0.27	17	60
55	1.15	0.20	23	50
80	1.1	0.13	30	40
			49	25
			58	20
			100	10
			150	5
			230	2
			300	1

Since the data on which this is based shows no effect whatsoever due to the enormous curtailments of publication volume in both World War I and World War II we can be sure that what we have here is pure obsolescence, unaffected by the size of the available literatures. The time variation is rather different from anything that had been previously proposed and follows a typically *S*-shaped logistic decline with the logarithm of time as the independent variable. I conjecture that the form might be $y = e^{-x}\ (1+x)$ where $x = \log_e$ (t/10 years) and that it probably arises from the usual differential equation defining the exponential growth of the literature, $dy/dt = Ky$ being modified to include a term on the right of the second degree, $dy/dt = Ky - Ly^2$.

At any event, what seems to happen in the process of obsolescence as we now can see it, is that during the first several years after publication the utility of relative citability of a paper declines only very slowly and parabolically in the logarithm of years elapsed. Even after *a century* the chance of citation has decreased by only a factor of ten. Most citations are to recent papers because most papers *are* recent, and it is dubious if there is anything of an immediacy effect due to rapid short-range obsolescence as I once conjectured. Obsolescence would seem to be an essentially long-range phenomena, akin to the effect of finite lifetime of authors in curtailing publication productivity. Since we now have the empirical values of G for papers of various ages, one may use the obsolescence data to derive from the time spectrum of citation the volume of publication that must exist at the various dates in the past century. One may also derive the citation frequency distributions for papers of varying age. It is evident immediately that highly cited papers will fall off more slowly with elapsed time than less cited papers.

• Philosophical Epilogue

I cannot conclude this first paper on the Cumulative Advantage Distribution without some remarks on its conceptual significance. The surface has only been scratched and doubtless the application of this theory will raise as many problems as are solved and demand much more empirical testing and rigorous statistical mathematics in expression. What intrigues me most is that this new underlying theory which seems to pull together so many diverse phenomena and qualitative laws into good quantitative predictability does so in a way that has an element of causal asymmetry. One would have supposed, for example, that the process of citation depended equally on the cited paper and the citing paper. It is with this in mind that some recent investigators (*35*) have been taking long hard looks at the apparent caprice with which a paper may or may not decide to cite previous work and may even make token or ritualistic citations.

In this theory, it would appear that the course of future citation successes is determined statistically by

the past history of the cited paper; and so one is driven to suppose that citations are generated by a pull mechanism from previous citation rather than from a push mechanism of the papers that do the citing. We even derive a relationship which goes some of the way to explaining the average number of references per paper as a consequence of the success distribution of the already existing corpus of literature, without there being much possible allowance for what such a reference actually implies.

It seems to me that the injection of some version of an underlying theory of this nature goes a long way towards solving the problem of what it is that we have been measuring and counting in bibliometric research and the other social science fields where cumulative advantage appears to operate.

Acknowledgments

I wish to record my personal gratitude to the Office of Science Information Service of the National Science Foundation for their friendly counseling over many years, and to Eugene Garfield and his staff at the Institute for Scientific Information for extensive and freely given cooperation in all matters pertaining to citation indexing which has made possible these interesting investigations. I thank also Mark d. Price for computing the table of the Beta Function, I. Richard Savage for some discussions of statistical matters, Gabriel Pinski and Christopher Anagnostakis for detecting and correcting an error in my initial proof of the distribution law, and Belver C. Griffith for a seemingly innocent remark which led me from an erroneous draft reply to a Garfield editorial towards the present work, which brings to a head this long series of researches.

References

1. **Simon, H.A.** 1957. *Models of Man: Social and Rational.* Chapter 9: On a Class of Skew Distribution Functions. New York: John Wiley & Sons, 1957.
2. **Feller, W.** *An Introduction to Probability Theory and Its Applications.* (3rd ed.) New York: John Wiley & Sons, I (1968), II (1966).
3. **Bliss, C.I.; Fisher, R.A.** 1953. "Fitting the Negative Binomial Distribution to Biological Data and Note on the Efficient Fitting of the Negative Binomial." *Biometrics.* The Biometric Society. 1953; 9(2): 176-200.
4. **Bliss, C.I.** "The Analysis of Insect Counts as Negative Binomial Distributions." *Proceedings of the Tenth International Congress of Entomology.* 2: 1015-1032 (1956 [1958]).
5. **Williamson, E.; Bretherton, M.H.** 1963. *Tables of the Negative Binomial Probability Distribution.* New York: John Wiley & Sons, 1963.
6. **Yule, G.U.** 1924. "A Mathematical Theory of Evolution, based on the Conclusions of Dr. J.C. Willis, F.R.S." *Philosophical Transactions of the Royal Society* 1924 B. 213: 21-87.
7. **Hill, B.** 1970. "Zipf's Law and Prior Distributions for the Composition of a Population." *Journal of the American Statistical Association.* 1970; 65 (331): 1220-1232.
8. **Hill, B.** 1974. "Rank Frequency Forms of Zipf's Law." *Journal of the American Statistical Association.* 1974; 69 (348): 1017-1026.
9. **Crowley, C.J.** 1975. "The Distribution of Citations to Scientific Papers: A Model." Presented at Midwest Sociological Society meeting, Chicago, April 1975 (unpublished).
10. **Hill, B.; Woodroofe, M.** 1975. "Stronger Forms of Zipf's Law." *Journal of the American Statistical Association.* 1975; 70 (349): 212-219.
11. **Woodroofe, M.; Hill, B.** 1975. "On Zipf's Law." *Journal of Applied Probability.* 1975; 12: 425-434.
12. **Sichel, H.S.** 1975. "On a Distribution Law for Word Frequencies." *Journal of the American Statistical Association.* 1975; 70 (352, part 1): 542-547.
13. **Cole, J.R.; Cole, S.** 1973. *Social Stratification in Science.* Chicago, IL: University of Chicago Press, 1973.
14. **Fairthorne, R.A.** 1969. "Progress in Documentation: Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction." *Journal of Documentation.* 1969; 25 (4): 319-343.
15. **Narin, F.** 1976. *Evaluation Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity.* Cherry Hill, NJ: Computer Horizons, 1976.
16. **Brookes, B.C.** 1969. "Bradford's Law and the Bibliography of Science." *Nature.* 1969; 224 (5223): 953-956.
17. **Naranan, S.** 1970. "Bradford's Law of Bibliography of Science: an Interpretation." *Nature.* 1970; 227: 631-632.
18. **Allison, P.D.; Price, D.de S.; Griffith, B.C.; Moravcsik, M.J.; Stewart, J.A.** 1976. "Lotka's Law: A Problem in its Interpretation and Application." *Social Studies of Science.* 1976; 6: 269-276.
19. **Groos, O.V.** 1967. Bradford's Law and the Keenan-Atherton Data." *American Documentation.* 1967; 18: 46.
20. **Barr, K.P.** 1967. "Estimates of the Number of Currently Available Scientific and Technical Periodicals." *Journal of Documentation.* 1967; 23 (No. 2): 110-116.
21. **Price, D. de S.** 1971. "Some Remarks on Elitism in Information and the Invisible College Phenomenon in Science." *Journal of the American Society for Information Science.* 1971; 22 (2): 74-75.
22. **Shockley, W.** 1957. "On the Statistics of Individual Variations of Productivity in Research Laboratories." *Proceedings of the Institute of Radio Engineers.* 1957; 45: 279-290, 1409-1410.
23. **Zener, C.** 1968. "An Analysis of Scientific Productivity." *Proceedings of the National Academy of Sciences.* 1968; 59 (4): 1078-1081.
24. **Lotka, A.J.** 1926. "The Frequency Distribution of Scientific Productivity." *Journal of the Washington Academy of Sciences.* 1926; 16 (12): 317-323.
25. **Price, D.de S.** 1963. *Little Science, Big Science.* New York: Columbia University Press. 1963.
26. **Price, D.de S.; Gürsey, S.** 1976. "Studies in Scientometrics. Part I: Transience and Continuance in Scientific Author-

ship." *International Forum on Information and Documentation.* International Federation for Documentation Proceedings. 1976; 1 (2): 17-24.
27. **Price, D.de S.** 1965. "Networks of Scientific Papers." *Science.* 1965; 149 (3683); 510-515.
28. **Yermish, I.** Private communication. 1976, March 17.
29. **Crowley, C.J.** 1975. "The Distribution of Citations to Scientific Papers: A Model." paper presented at Midwest Sociological Society meeting, Chicago. April 1975 (unpublished).
30. **Brookes, B.C.** 1975. "A Sampling Theorem for Finite Discrete Distributions." *Journal of Documentation.* 1975; 31 (1): 26-35.
31. **Garfield, E.** 1974. "Selecting the All-Time Citation Classics. Here Are the Fifty Most Cited Papers for 1961-1972." *Current Contents.* Editorial, (2): 1974 January 9.
32. **Garfield, E.** 1976. "Is the Ratio Between Number of Citations & Publications Cited a True Constant." *Current Contents,* Editorial, Current Comments, (6) 1976 February 9.
33. **Price, D.de S.; Gürsey, S.** 1976. "The Relation between Source Author and Cited Author Populations." *International Forum on Information and Documentation.* International Federation for Documentation. 1976; 1 (3): 19-22.
34. **Gomperts, M.C.** 1968. "The Law of Constant Citation for Scientific Literature." *Journal of Documentation.* 1968; 24 (2): 113-117.
35. **Moravcsik, M.J.; Murugesan, P.** 1975. "Some Results on the Function and Quality of Citations." *Social Studies of Science.* 1975; 5: 86-92.