Pergamon

Information Processing & Management, Vol 30, No. 4, pp. 535-556, 1994

0306-4573/94 $6.00 + .00

0306-4573(93)E0003-7

THE SIMON-YULE APPROACH TO BIBLIOMETRIC MODELING

YE-SHO CHEN
Department of Quantitative Business Analysis,
Louisiana State University, Baton Rouge, LA 70803, U.S.A.

P. PETE CHONG
Department of Operations and Information Systems, School of Business Administration,
Gonzaga University, WA 99258-0001, U.S.A.

and

MORGAN Y. TONG
Department of Quantitative Business Analysis,
Louisiana State University, Baton Rouge, LA 70803, U.S.A.

(*Received* 11 *May* 1993; *accepted in final form* 13 *October* 1993)

Abstract—Using an index approach to take into account the scattering pattern of the observed values, Chen and Leimkuhler showed that the three well-known bibliometric distributions (i.e., Lotka's law of scientific productivity, Bradford's law of bibliographic scattering, and Zipf's law of word frequency) are equivalent. Furthermore, Chen showed that Lotka's law can be derived from a generating mechanism (the Simon-Yule Model) proposed by Herbert A. Simon. In this paper, we use a simulation algorithm based on the Simon-Yule model to conduct computational experimentation on these three laws. The results indicate that the probability of a new entry (α), be it constant or decreasing, determines the characteristics of all three distributions.

I. INTRODUCTION

Three bibliometric distributions are well known in the information science community; they are Lotka's law of scientific productivity (Lotka, 1926), Bradford's law of bibliographic scattering (Bradford, 1934), and Zipf's law of word frequency (Zipf, 1949). Descriptive arguments about the equivalence of these three laws have been reported in the literature. To provide a common functional relationship among the three bibliometric distributions, Chen and Leimkuhler (1986) proposed a more rigorous approach that, by means of an index, takes explicit account of the sequence of observed variable values. The same index approach was used to explain the droop phenomenon of Bradford's law (Chen & Leimkuhler, 1987a), the concave abnormality of Zipf's law (Chen & Leimkuhler, 1987b), and Booth's law of low-frequency words (Chen & Leimkuhler, 1990)—a dual phenomenon of Zipf's Law.

Through this index approach, the study of bibliometric distributions boils down to the stochastic modeling of Lotka's law. The development of stochastic bibliometric models faces considerable problems in verifying, validating, and experimenting with these models (Leimkuhler, 1988). Simon (1977) discussed these problems at length, and he proposed a more constructive approach that consists of the following five steps: (1) Begin with raw data, not theories; (2) draw simple generalizations from striking features in data; (3) find limiting conditions by manipulating the variables; (4) devise simple mechanisms to explain steps 2 and 3; and (5) propose explanatory theories that go beyond step 4 and make experiments. These five steps were adopted by Chen (1989) to show that through the index approach, Lotka's law can be derived from the Simon-Yule model—a generating mechanism proposed by Simon (1955).

Because the conventional analytical methods can only derive the "average behavior" of the distributions, Leimkuhler (1988) suggested the use of computational experimenta-

tion for bibliometric modeling, since it enables us to study the distributions under "extreme conditions" (Neuts, 1986a, 1986b). The aim of the paper is to investigate, via simulation, the relationships between the parameters of the Simon-Yule model and the shapes of the three bibliometric distributions. Specifically, computational experimentation based on two versions of the Simon-Yule model are conducted on Lotka's law, Bradford's law, and Zipf's law. Through changing the parameters of the Simon-Yule model, we are able to explain the regularities and anomalies of the three laws. Furthermore, a time dimension is included in the computational experimentation, which provides a valuable insight into the stability of the empirical laws.

This paper is organized as follows. Section 2 provides an overview of the three bibliometric distributions mentioned earlier. Section 3 describes the Simon-Yule model and the simulation algorithm derived from it to generate the three distributions; we also argue that the computational experimentation is the most promising approach in this instance. Sections 4 to 6 are the computational results of these three distributions. Additional effects of changing parameters are discussed in Section 7; and finally, Section 8 is the conclusion.

2. ANALYSIS OF THE BIBLIOMETRIC DISTRIBUTIONS: AN INDEX APPROACH

2.1 *The bibliometric distributions*

In his 1926 paper, Lotka examined patterns of scientific productivity among chemists. He discovered that, for some positive constant a, the number of chemists who published n papers was approximately a/n^2, or

$$f(n) = an^{-2}, \quad n = 1,2,3,\ldots.$$

Letting $F(n) = \sum_{i=n}^{\infty} f(i)$ be the number of authors who published n or more papers, then a frequently used alternative form of Lotka's law is

$$F(n) = a\int_n^{\infty} \frac{1}{x^2}\,dx = an^{-1}, \quad n = 1,2,3,\ldots \tag{1}$$

Bradford discovered in 1934 that if a comprehensive search on one particular topic is carried out for a period of time, and journals are arranged in descending order according to the number of articles found in them, the sources can be divided into a nucleus of journals and several zones containing the same number of articles as the nucleus, and the number of journals in the nucleus and succeeding zones will be

$$1 : j : j^2 : \ldots, \tag{2}$$

for some constant j.

In his 1949 book, Zipf stated that "if one takes the words making up an extended body of text and ranks them by frequency of occurrence, then the rank r multiplied by its frequency of occurrence, $g(r)$, will be approximately constant." In symbolic form,

$$g(r) = br^{-1}, \quad r = 1,2,3,\ldots \tag{3}$$

where b is a positive constant.

As Chen and Leimkuhler (1986) pointed out, each of these three laws studies a particular arrangement of two groups: the observation and the class. Lotka's law relates the observation (the papers) and the class (the authors) by their frequency-size relationship. Bradford's law relates the observation (the papers) and the class (the journals) by the cumulative-frequency-log-rank approach. Zipf's law focuses on the observation (the word occurrences) and the class (the words) by their frequency-rank relationship. More recently, the terminology "item" and "source" are used more often (e.g., Egghe & Rousseau, 1989, 1990)

than the observation and the class. As such, we use the term item-source in the rest of this paper as the observation-class relationship.

2.2 *The index approach and its contributions*

Chen and Leimkuhler (1986) introduced the notion of an index $i = 1,2,\ldots,m$, to take the scattering pattern of the observed values of the bibliometric distributions into account. In terms of items and sources, we define

n_i = the ith different observed value of n items a source has such that $n_{i+1} > n_i$,
$f(n_i)$ = the number of sources with n_i items,
$F(n_i)$ = the number of sources that have no fewer than n_i items,
r_i = the ith observed rank of a source where rank depends on the frequency of its items,
$g(r_i)$ = the number of items of a source with rank r_i, and
$G(r_i)$ = the number of items of sources with rank not greater than r_i.

Three contributions of the indexed approach can be identified. First, Chen and Leimkuhler (1986) showed that the three empirical laws are mathematically equivalent. Using notations defined above, for $i = 1,2,\ldots,m$,

$$F(n_i) = an_i^b - c \tag{4}$$

$$\text{iff } G(r_i) = d\sum_{k=1}^{i}[r_k^e(r_k - r_{k-1})] \tag{5}$$

and

$$\text{iff } g(r_i) = d(r_i + c)^e, \tag{6}$$

where a,b,c,d,e are constants and $a,d > 0$; $b,e < 0$; $be = 1$, $ad^b = 1$. Equations (4), (5), and (6), without the index notations, are general formulations of Lotka's law, Bradford's law, and Mandelbrot-Zipf's law (Mandelbrot, 1953), respectively.

Second, the indexes $i = 1,2,\ldots,m$, can be divided into three regions: where i is small, where i close to m, and otherwise. For small i, $n_i = i$. For i close to m, $f(n_i) = 1$. Let i_ℓ be the maximum i such that $n_i = i$ and let i_u be the minimum i such that $f(n_i) = 1$ and $f(n_{i-1}) \neq 1$. Then we have the following three important properties:

1. $n_i = i$, $\quad 1 \leq i \leq i_\ell$,
2. $n_i \simeq i$ and $f(n_i) \simeq 1$, $\quad i_\ell + 1 \leq i \leq i_u - 1$,

and

3. $f(n_i) = 1$, $\quad i_u \leq i \leq m$,

where $\simeq$ means approximately equal. These three properties enable us to describe the droop phenomenon of Bradford's law (Chen & Leimkuhler, 1987a) and the concave abnormality of Zipf's law (Chen & Leimkuhler 1987b).

Third, Zipf's law focuses mainly on words of high frequency. In contrast, the formulation of Booth's law (Booth, 1967) was motivated by two remarkable phenomena associated with words that rarely occurred. Letting $f(n)$ be the number of words appearing n times each in a literary text and T be the total number of different words in the same text, then $f(1)/T \simeq 0.5$, $f(2)/f(1) \simeq 0.33$, $f(3)/f(1) \simeq 0.17$, $f(4)/f(1) \simeq 0.10$, and $f(5)/f(1) \simeq 0.07$. Chen and Leimkuhler (1990) showed that these equations can also be derived through eqn (4) (i.e., the indexed version of Lotka's law).

3. GENERATING BIBLIOMETRIC DISTRIBUTIONS: THE SIMON-YULE APPROACH

After analyzing the bibliometric distributions in Section 2, the logical next step is to study the indexed version of Lotka's law (eqn (4)). Chen (1989) adopted Simon's (1977)

five-step scientific modeling process (see Section 1) to model Lotka's law. The modeling process can be briefly summarized as follows: (a) examining empirical data of Lotka's law using the index approach, (b) some striking features of the data related to eqn (4) are observed and generalized, (c) influential variables associated with the data are identified, (d) a simple generating mechanism proposed by Simon (1955) is used to derive eqn (4), and (e) need for further refinement of the simple mechanism is discussed. The simple mechanism, called the Simon-Yule model, and its further refinement are briefly reviewed below.

3.1 *The Simon-Yule Model of Lotka's Law*

According to Simon (1955), the way we select things to use can be described as a stochastic process; furthermore, it is a twofold process that consists of imitation and association. People select what to use according to what they have used before and what others are using (imitation), as well as what they have just recently used (association). Although Simon's original subject matter was the text generation, in terms of items and sources, his assumptions can be generalized to be:

1. there is a constant probability, α, and the $(t+1)$st item will be a source (i.e., a source that has not been used in the first t items); and
2. the probability that the source corresponding to the $(t+1)$st item used has been used n times before is proportional to $n \cdot f(n,t)$, where $f(n,t)$ is the number of distinct sources used exactly n times each in the first t items.

Although the Simon-Yule model provides a sound generating mechanism to explain the general form of Lotka's law in eqn (4) (Chen, 1989), several questions need to be addressed. For example, (a) what is the impact on the distribution when the size of the probability, α, is varied? (b) Is there any relationship among m (the index), n_m (the largest n), and α? (c) Is the average items per source (i.e., the total number of items over the total number of sources) a function of α?

3.2 *Further refinement of the Simon-Yule Model*

As noted by Simon and Van Wormer (1963), the above-mentioned model is only a first approximation of the reality. They further refined the model by modifying the α in the first assumption from constant probability to a decreasing function of the total number of items. That is, there is a decreasing probability function $\alpha(t)$, $0 \leq \alpha(t) \leq 1$, that the $(t+1)$st item used is from a source not used previously. An immediate question is whether a decreasing function will provide different answers to the three questions listed above than constant function does.

3.3 *The need for computational experimentation*

In his paper on bibliometric modeling, Leimkuhler (1988) argued that the use of computational experimentation is necessary in studying bibliometrics and their application to information system design and problem solving. Computational experimentation allows researchers to go beyond the analytical methods to examine in detail as the assumptions are relaxed (Simon & Van Wormer, 1963). The two versions of the Simon-Yule model shown in the previous subsections can be easily programmed on a computer to simulate empirical data $(n_i, f(n_i))$, $i = 1,2,\ldots,m$, which exhibit Lotka's law (Chen, 1989).

Let N be the total number of items and $f(j,t)$ be the number of distinct sources used exactly j times each in the first t items. The simulation algorithm proposed by Simon (Simon & Van Wormer, 1963) consists of two steps:

Step 1. For each t $(1 \leq t \leq N)$, a random number, a, is generated from the rectangular distribution with range 0 to 1. If $a \leq \alpha(t)$, $f(1,t) = f(1,t-1) + 1$ (i.e., a new source is added to the set of previously used sources); otherwise go to step 2.

Step 2. A random number, b, is drawn from the rectangular distribution with range $1 \leq b \leq t$. Starting with $j = 1$, the cumulant of $j \cdot f(j,t-1)$ is computed, and

compared with b until an n is found such that $\sum_{j=1}^{n} jf(j,t-1) \geq b$. Then $f(n,t) = f(n,t-1) - 1$; and $f(n+1,t) = f(n+1,t-1) + 1$. This is equivalent to saying that this tth item used belongs to the group of sources that have been used n times previously, and now it is moved to another group where every source has been used $n+1$ times.

The simulation data are obtained from a computer program written in *Turbo Pascal* for an 80386 personal computer. To start the simulation program, the initial conditions $f(n,0)$, $n = 1,\ldots,N$, shall be provided. Since moderate changes in the initial conditions do not appear to affect the equilibrium distributions of Lotka's law, we set the initial conditions with $f(1,0) = 3$, and $f(n,0) = 0$ for $n = 2,\ldots,N$. In other words, the first three usages involve three different items. The simulations are carried out with N ranging from 1,000 to 30,000. On the other dimension, α also varied. For constant α, it ranges from 0.1 to 0.9 and 0.1 increment plus the two extreme conditions of $\alpha = 0.01$ and $\alpha = 0.99$. For the decreasing function, we use $\alpha(R) = A/\ln(R)$, where A ranges from 1.00 to 2.00 with the increment of 0.25, and $R = 1,2,\ldots,N$. The simulation data generated from Simon's algorithm are then entered into a Lotus 1-2-3 spreadsheet to facilitate transformation. Graphs are also created from the transformed data for further analyses.

4. COMPUTATIONAL EXPERIMENTATION OF LOTKA'S LAW

4.1 *Constant entry rate*

Figure 1 is one example of the results of simulations using Lotka's law with $N = 20{,}000$ and $\alpha = 0.10$ and 0.90. A full graph reveals little information, since the curves are crowded along the axes. Thus, we magnify and focus only on the part closest to the origin to decipher the results. Figure 1 shows that, in general, large α tends to generate a smaller n_m, the frequency of usage for the most used source, and it also reduces the curve to a near vertical line approaching the y axis. In other words, higher α shows a more evenly distrib-

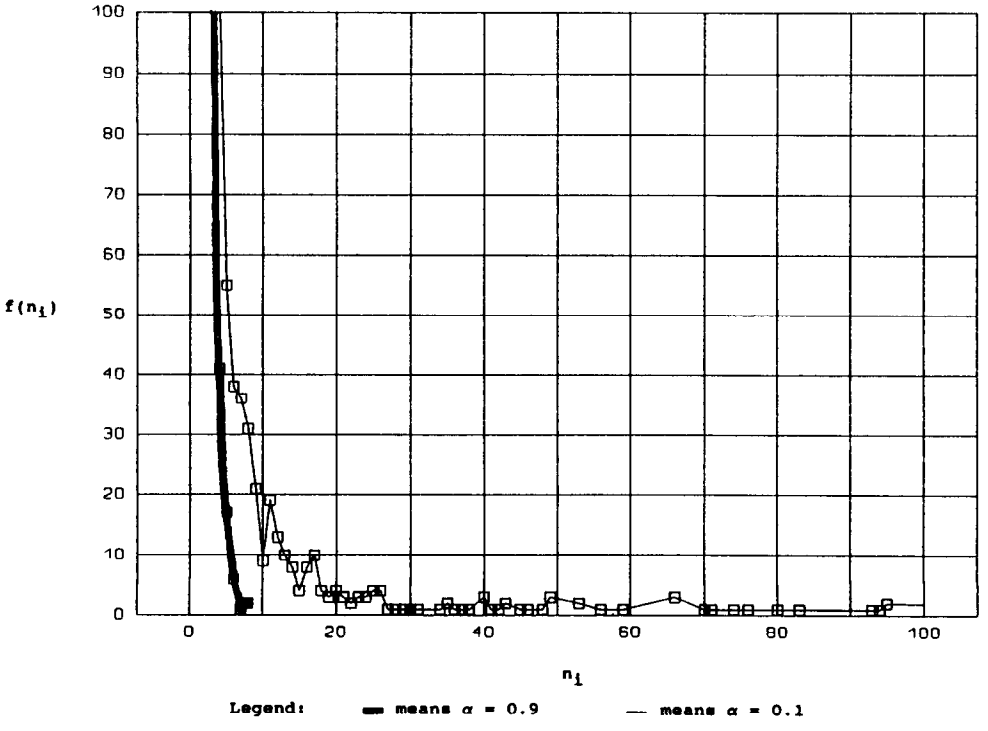

Fig. 1. Lotka's law with constant α ($N = 20{,}000$).

uted usage of sources. In terms of the three regions described in Section 2.2, high α decreases and low α increases the region 3 where $f(n) = 1$. In fact, when $\alpha \simeq 0.90$, $n_m = m$ and there is no excessive region 3. For example, when $N = 20{,}000$ $n_m = m = 8$ (i.e., the most used sources are used eight times each) (see Table 1). In contrast, when $\alpha = 0.01$, the frequency of usage for the most used source is 10,578—a highly concentrated result.

In light of the difficulty in describing different curvatures of these graphs, we need some measurement to show the relationship among parameters. Since Fig. 1 shows that a large α seems to have less area under its curve and a small α implies large area, we define the parameter Area to be the area under the curve formed by $\{(n_i, f(n_i)), i = 1,2,\ldots,m\}$; that is,

$$\text{Area} = \tfrac{1}{2}[(f(n_1) + f(n_2))(n_2 - n_1) + (f(n_2) + f(n_3))(n_3 - n_2) + \ldots + (f(n_{m-1}) + f(n_m))(n_m - n_{m-1})]. \quad (7)$$

The parameter Area serves as a measurement of the concentration of the usage of the sources. Note that more classical and similar concentration measurements do exist (e.g., Gini's index). The use of the parameter Area in eqn (7) (and later in eqns (8) and (9)) is due to its fitness for the empirical law being studied.

If we denote this nominal Area under the Lotka's curve to be Area_1, then from Table 1 we can see that Area_1 increases linearly with respect to the size of N when α is held equal. For example, at $\alpha = 0.01$, when $N = 30{,}000$ $\text{Area}_1 = 15943$, or approximately 30 times the Area_1 of 536 when $N = 1{,}000$. For this reason, we arbitrarily selected $N = 20{,}000$ to be representative in most of our discussions. Since a larger N automatically increases Area_1, Area_1 is adjusted and expressed as a percentage of the corresponding $n_m f(1)$—the rectangular area with the two extreme points as the corners. The resulting fraction is denoted A_L and listed in Table 1. Figure 2 indicates how A_L varies at different levels of α and N. Figure 2 also shows that regardless the magnitude of N, A_L increases along with α, and finally A_L converges to approximately 16.97% (when $\alpha = 0.99$) for all N. The reason for this convergence at high level of α would be a good topic for future research. Note also that at the other extreme condition, $\alpha = 0.01$, A_L also defies this general pattern of positive correlation with α.

The ratios according to Booth's law are also included in this table. Note that when $\alpha \simeq 0.20$, the ratios approximate those expressed in Section 2.2. Note also that the discrepancy between m (index) and n_m (the largest n in the index system) in Table 1 indicates the nature of scattering observed values in bibliometric distributions.

4.2 *Decreasing entry rate*

Figure 3 shows that when A varies from 1.0 to 2.0, the curve moves away from the origin. However, results in Table 2 also indicate that although Area_1 increases when α increases, the rate of increase is not as significant as in the case of constant α. Furthermore, when N is large, the rate of increase is even smaller. This is understandable, since the decreasing function eventually generates an α that is too small for A to make a significant difference. However, note how well the Booth's ratios comply with the theoretical figures in Section 2.2! For instance, at $A = 1.5$ and $N = 20{,}000$, the ratios are approximately 0.50, 0.33, 0.17, 0.11, and 0.06.

Figure 4 shows that A_L, as defined in Section 4.1, increases as A (and therefore α) increases in all levels of N. A_L increases following a smooth slope for all N except that of $N = 1{,}000$. It is apparent that the decreasing function has not yet had a chance to remove the volatility at that level of N. The decreasing A_L with respect to N is caused largely by the much faster increase in $n_m f(1)$.

5. COMPUTATIONAL EXPERIMENTATION OF BRADFORD'S LAW

We began the analysis of Bradford's law by calculating the Area, denoted as Area_B, using the following formula, with the same notations defined in Section 2.2:

Table 1. Simulation results of Lotka's law with constant α

α	N (000)	m	n_m	T	$f(1)/T$	$f(2)/f(1)$	$f(3)/f(1)$	$f(4)/f(1)$	$f(5)/f(1)$	Area_L	A_L
0.01	1	6	534	9	0.4444	0.0000	0.2500	0.0000	0.0000	536	0.2509
0.10	1	16	371	88	0.5341	0.2128	0.1702	0.1915	0.0638	420	0.0241
0.20	1	22	175	183	0.4918	0.4556	0.1556	0.0556	0.0556	292	0.0185
0.30	1	21	91	302	0.5993	0.2099	0.1381	0.1105	0.0497	282	0.0171
0.40	1	18	44	401	0.6060	0.2510	0.1728	0.0658	0.0412	316	0.0295
0.50	1	17	35	496	0.6431	0.2978	0.0909	0.0596	0.0251	356	0.0319
0.60	1	13	16	611	0.7201	0.2114	0.0818	0.0341	0.0136	394	0.0559
0.70	1	10	13	716	0.7765	0.1871	0.0486	0.0306	0.0018	441	0.0609
0.80	1	6	8	799	0.8260	0.1530	0.0364	0.0152	0.0030	472	0.0894
0.90	1	5	5	906	0.9205	0.1367	0.0360	0.0144	0.0120	488	0.1170
0.99	1	3	3	987	0.9878	0.0113	0.0010	0.0000	0.0000	499	0.1706
0.01	5	13	2624	44	0.5682	0.2000	0.0800	0.1200	0.0400	2642	0.0403
0.10	5	34	1618	474	0.5359	0.2913	0.1654	0.0748	0.0591	1935	0.0047
0.20	5	43	642	990	0.5424	0.3240	0.1620	0.0950	0.0447	1331	0.0039
0.30	5	40	303	1523	0.5896	0.3129	0.1102	0.0735	0.0367	1339	0.0049
0.40	5	39	109	2014	0.6346	0.2551	0.1072	0.0689	0.0352	1450	0.0104
0.50	5	27	67	2487	0.6639	0.2489	0.1127	0.0424	0.0279	1705	0.0154
0.60	5	21	26	2958	0.7093	0.2288	0.0791	0.0381	0.0186	1918	0.0352
0.70	5	15	19	3450	0.7609	0.1996	0.0530	0.0232	0.0175	2142	0.0429
0.80	5	9	12	3954	0.8283	0.1450	0.0345	0.0168	0.0046	2321	0.0590
0.90	5	7	7	4492	0.9103	0.0792	0.0147	0.0037	0.0005	2447	0.0855
0.99	5	3	3	4957	0.9917	0.0079	0.0004	0.0000	0.0000	2498	0.1694
0.01	10	16	5235	105	0.5143	0.3333	0.1481	0.0741	0.0926	5300	0.0187
0.10	10	46	3015	998	0.5411	0.3019	0.1204	0.0963	0.0667	3704	0.0023
0.20	10	55	1120	2018	0.5476	0.3113	0.1620	0.0887	0.0516	2545	0.0021
0.30	10	52	506	3036	0.5827	0.3041	0.1357	0.0820	0.0447	2626	0.0029
0.40	10	49	174	3985	0.6211	0.2764	0.1329	0.0549	0.0303	2879	0.0067
0.50	10	38	94	4915	0.6596	0.2619	0.0956	0.0506	0.0312	3353	0.0110
0.60	10	27	36	5933	0.7106	0.2265	0.0804	0.0372	0.0209	3838	0.0253
0.70	10	19	25	6934	0.7651	0.1919	0.0586	0.0234	0.0124	4287	0.0323
0.80	10	10	12	7939	0.8293	0.1472	0.0328	0.0150	0.0050	4649	0.0588
0.90	10	7	7	9008	0.9112	0.0803	0.0125	0.0035	0.0007	4904	0.0853
0.99	10	3	3	9918	0.9919	0.0079	0.0002	0.0000	0.0000	4998	0.1693
0.01	15	19	7917	149	0.4698	0.3429	0.2857	0.0857	0.0714	8034	0.0145
0.10	15	57	4383	1504	0.5259	0.3338	0.1517	0.0796	0.0544	5446	0.0016
0.20	15	65	1547	3039	0.5535	0.3002	0.1576	0.0779	0.0559	3703	0.0014
0.30	15	58	649	4538	0.5837	0.2990	0.1351	0.0774	0.0521	3838	0.0022
0.40	15	56	220	5930	0.6147	0.2914	0.1265	0.0642	0.0326	4285	0.0053
0.50	15	43	119	7388	0.6547	0.2756	0.0980	0.0476	0.0269	5056	0.0088
0.60	15	30	46	8917	0.7102	0.2291	0.0764	0.0403	0.0210	5777	0.0198
0.70	15	20	28	10482	0.7724	0.1821	0.0581	0.0240	0.0120	6446	0.0284
0.80	15	12	15	11954	0.8329	0.1443	0.0318	0.0131	0.0061	6980	0.0467
0.90	15	8	8	13491	0.9106	0.0801	0.0136	0.0031	0.0010	7348	0.0748
0.99	15	3	3	14871	0.9915	0.0085	0.0001	0.0000	0.0000	7498	0.1695
0.01	20	23	10578	205	0.4780	0.3878	0.1837	0.1327	0.0612	10713	0.0103
0.10	20	66	5722	1972	0.5198	0.3356	0.1483	0.1063	0.0537	7162	0.0012
0.20	20	74	1935	3991	0.5437	0.3217	0.1581	0.0820	0.0539	4780	0.0011
0.30	20	69	783	6023	0.5816	0.3060	0.1285	0.0748	0.0514	4997	0.0018
0.40	20	60	260	8011	0.6243	0.2729	0.1244	0.0630	0.0338	5719	0.0044
0.50	20	48	137	9916	0.6632	0.2576	0.0999	0.0465	0.0295	6719	0.0075
0.60	20	34	48	11899	0.7131	0.2209	0.0844	0.0346	0.0206	7674	0.0188
0.70	20	23	31	13963	0.7726	0.1813	0.0586	0.0227	0.0125	8578	0.0256
0.80	20	12	15	15936	0.8328	0.1435	0.0333	0.0129	0.0057	9309	0.0468
0.90	20	8	8	18015	0.9112	0.0804	0.0129	0.0025	0.0010	9806	0.0747
0.99	20	3	3	19816	0.9909	0.0091	0.0002	0.0000	0.0000	9997	0.1697
0.01	25	27	13216	246	0.5000	0.3008	0.2033	0.1057	0.0650	13376	0.0082
0.10	25	72	6961	2435	0.5170	0.3249	0.1581	0.1033	0.0627	8710	0.0010
0.20	25	81	2332	4993	0.5484	0.3112	0.1494	0.0829	0.0606	5883	0.0009
0.30	25	75	917	7539	0.5832	0.2986	0.1321	0.0771	0.0484	6202	0.0015
0.40	25	66	295	9992	0.6247	0.2695	0.1245	0.0628	0.0344	7114	0.0039
0.50	25	50	148	12417	0.6632	0.2590	0.0988	0.0488	0.0270	8404	0.0069
0.60	25	37	50	14892	0.7123	0.2252	0.0802	0.0347	0.0228	9606	0.0181
0.70	25	23	32	17470	0.7728	0.1810	0.0581	0.0233	0.0132	10731	0.0248
0.80	25	13	16	19958	0.8342	0.1412	0.0348	0.0121	0.0055	11637	0.0437
0.90	25	8	8	22509	0.9109	0.0803	0.0134	0.0027	0.0009	12256	0.0747
0.99	25	3	3	24770	0.9909	0.0090	0.0002	0.0000	0.0000	12496	0.1697

continued

Table 1. continued

α	N (000)	m	n_m	T	$f(1)/T$	$f(2)/f(1)$	$f(3)/f(1)$	$f(4)/f(1)$	$f(5)/f(1)$	Area_1	A_1
0.01	30	28	15740	304	0.5033	0.3203	0.2288	0.0784	0.0523	15943	0.0066
0 10	30	84	8154	2915	0.5252	0.2972	0.1450	0.1156	0.0614	10224	0.0008
0.20	30	84	2677	5988	0.5464	0.3139	0.1494	0.0889	0.0581	6964	0.0008
0.30	30	83	1031	9033	0.5818	0.2986	0.1376	0.0735	0.0464	7370	0.0014
0.40	30	68	327	11987	0.6243	0.2700	0.1235	0.0643	0.0339	8529	0.0035
0.50	30	55	161	14891	0.6619	0.2605	0.0989	0.0518	0.0264	10071	0.0063
0.60	30	38	51	17899	0.7122	0.2255	0.0813	0 0346	0.0223	11545	0.0178
0.70	30	25	32	21000	0.7723	0.1839	0.0572	0.0231	0 0118	12898	0.0249
0.80	30	13	16	23988	0.8348	0.1416	0.0334	0.0123	0.0055	13980	0.0436
0.90	30	8	8	27036	0.9112	0.0804	0.0131	0.0026	0.0009	14717	0.0747
0.99	30	3	3	29722	0.9908	0.0090	0.0002	0.0000	0 0000	14994	0.1697

$$\text{Area}_B = \tfrac{1}{2}[(G(r_2) + G(r_1))(\log r_2 - \log r_1) + (G(r_3) + G(r_2))(\log r_3 - \log r_2) + \ldots + (G(r_m) + G(r_{m-1})(\log r_m - \log r_{m-1})]. \quad (8)$$

Furthermore, we define $A_B = \text{Area}_B/[G(r_m)\log(r_m)]$, where the denominator is the largest area possible for the curve.

5.1 *Constant entry rate*

Using $N = 20{,}000$ as an example, Fig. 5 is the composite graph of four Bradford's curves with $\alpha = 0.01$, 0.20, 0.40, and 0.90. Note that the graph is the functional representation of what is usually called Leimkuhler' law (1967). Thus, Leimkuhler-Bradford's law/curve is used in Figs. 5 to 8.

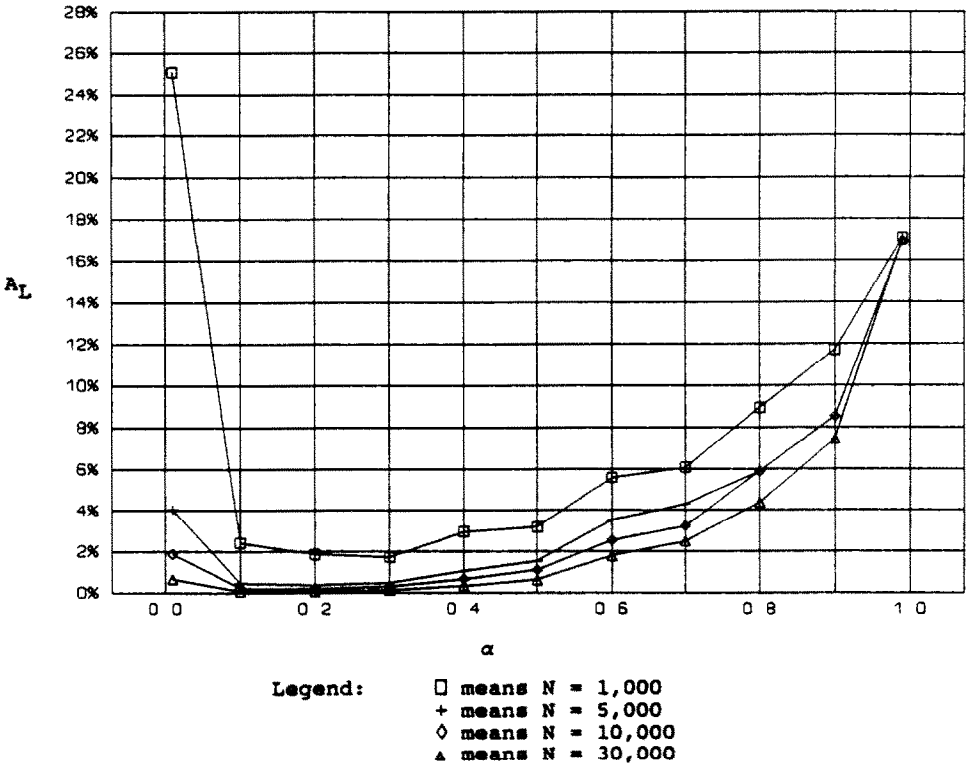

Fig. 2. Area under Lotka's curve (A_1) with constant α.

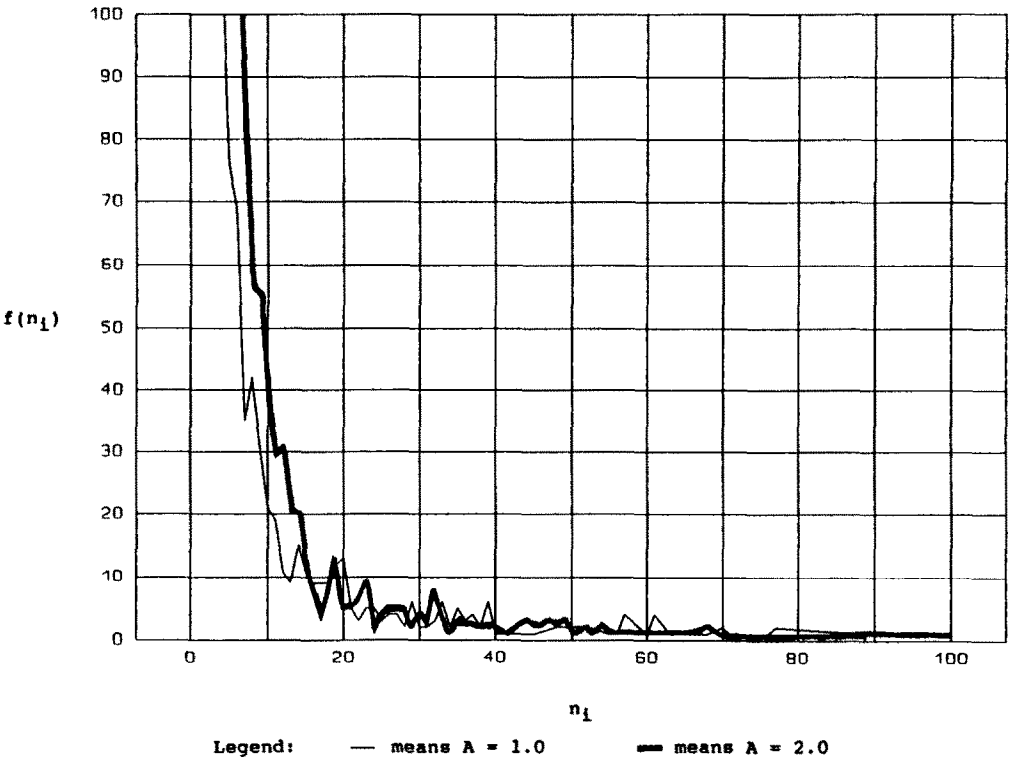

Fig. 3. Lotka's law with $\alpha(R) = A/\ln(R)$, $R = 1,2,\ldots,20{,}000$.

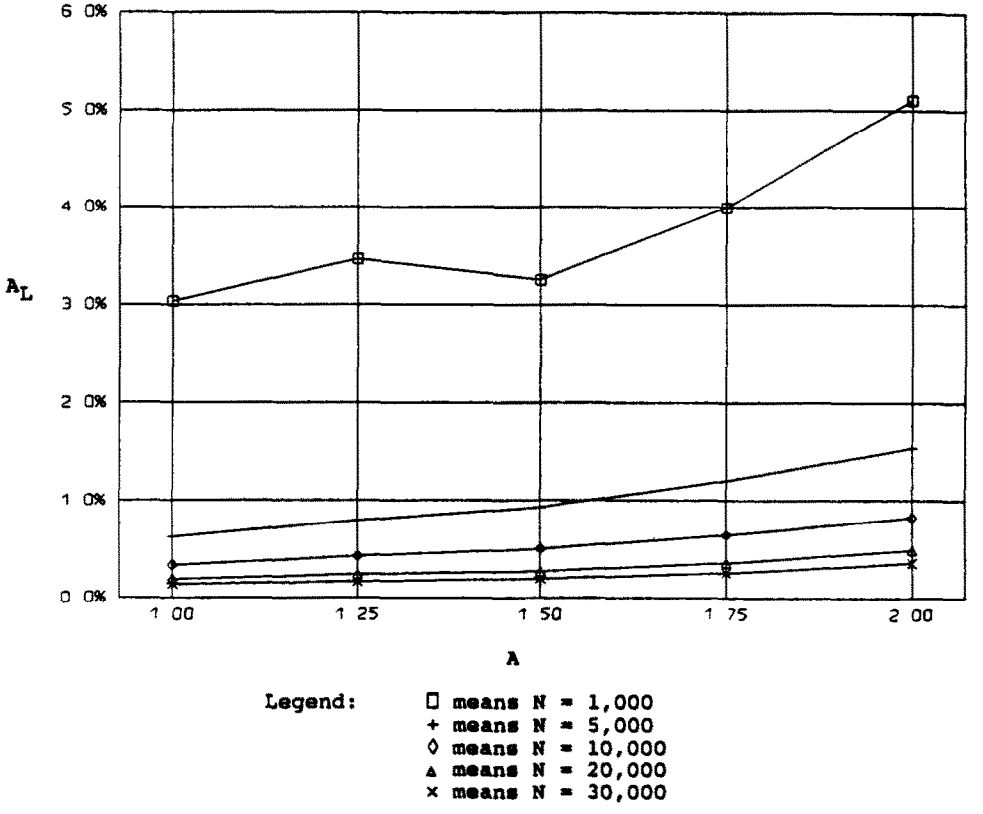

Fig. 4. Areas under Lotka's curve (A_L) with $\alpha(R) = A/\ln(R)$, $R = 1,2,\ldots,20{,}000$.

Table 2. Simulation results of Lotka's law, with $\alpha(R) = A/\ln(R)$ where $R = 1,2,\ldots,N$

A	N (000)	m	n_m	T	$f(1)/T$	$f(2)/f(1)$	$f(3)/f(1)$	$f(4)/f(1)$	$f(5)/f(1)$	Area_t	A_t
1.00	1	26	101	169	0.4142	0.6143	0.1571	0.1000	0.0429	214.5	0.0303
1.25	1	24	62	212	0.4481	0.4421	0.1895	0.0526	0.0421	204.0	0.0346
1.50	1	23	51	271	0.5129	0.3309	0.1079	0.1079	0.1079	230.5	0.0325
1.75	1	22	39	317	0.5205	0.3091	0.1394	0.1818	0.0727	257.0	0.0399
2.00	1	21	28	348	0.5316	0 2811	0.1676	0.1622	0.0595	264.5	0.0511
1.00	5	48	426	686	0.5044	0.2977	0.1590	0.0838	0.0694	917.0	0.0062
1.25	5	55	254	845	0.4935	0.3309	0.1511	0.1199	0.0959	838.0	0.0079
1.50	5	48	194	1033	0.4976	0.3346	0.1868	0.1109	0.0603	928.0	0.0093
1.75	5	49	133	1219	0.5086	0.3387	0.1935	0.0952	0.0645	994.0	0.0121
2.00	5	42	97	1367	0.5267	0.3417	0.1528	0.0931	0.0569	1076.0	0.0154
1.00	10	68	794	1262	0.4992	0.2952	0.1714	0.0984	0.0810	1676.5	0.0034
1.25	10	68	476	1576	0.5006	0.3105	0.1736	0.1065	0.0659	1606.0	0.0043
1.50	10	68	365	1889	0.4971	0.3365	0.1715	0.1076	0.0756	1724.5	0.0050
1.75	10	67	252	2235	0.5087	0.3369	0.1803	0.0915	0.0730	1853.5	0 0065
2.00	10	62	182	2532	0.5257	0.3366	0.1630	0.0924	0.0594	1996.0	0 0082
1.00	15	80	1143	1800	0.4894	0.3326	0.1657	0.0988	0 0795	2427.0	0.0024
1.25	15	80	675	2227	0.4926	0.3254	0.1778	0.1112	0 0720	2298.5	0.0031
1.50	15	80	515	2700	0.5056	0.3216	0.1692	0.1070	0.0571	2471.5	0.0035
1.75	15	80	356	3198	0.5181	0.3132	0.1768	0 0863	0.0670	2654.0	0.0045
2.00	15	77	248	3600	0.5208	0.3419	0.1552	0.1003	0.0624	2844.0	0.0061
1.00	20	84	1500	2280	0.4816	0 3597	0.1585	0.1056	0.0692	3176.5	0 0019
1.25	20	96	879	2861	0.4939	0.3390	0.1599	0.1125	0.0672	2964.0	0.0024
1.50	20	95	673	3445	0.5030	0.3289	0.1679	0.1062	0.0600	3169.0	0.0027
1.75	20	88	460	4069	0.5092	0.3369	0.1747	0.0936	0.0565	3435.5	0.0036
2.00	20	89	314	4586	0.5118	0.3524	0.1670	0.0950	0 0609	4642.0	0.0049
1.00	25	95	1854	2747	0.4816	0.3492	0.1602	0.1156	0.0582	3883.0	0.0016
1.25	25	100	1074	3454	0.4939	0.3353	0.1530	0.1184	0.0645	3604.5	0.0020
1.50	25	102	800	4174	0.5026	0.3308	0.1540	0.1115	0.0686	3849.5	0.0023
1.75	25	104	539	4920	0.5106	0.3229	0.1684	0.0999	0.0685	4109.0	0.0030
2.00	25	100	362	5583	0.5182	0.3246	0.1701	0.0954	0.0667	4409.5	0.0042
1.00	30	101	2192	3220	0.4904	0.3097	0.1659	0.1203	0.0665	4657.0	0.0013
1.25	30	111	1260	4047	0.4977	0.3133	0.1708	0 1087	0.0665	4210.0	0.0017
1.50	30	111	955	4895	0.5046	0.3186	0.1688	0.0980	0.0745	4590.0	0.0019
1.75	30	110	641	5764	0.5095	0.3201	0.1709	0.1042	0.0678	4858.0	0.0026
2.00	30	109	432	6545	0 5141	0.3337	0.1664	0.1004	0 0627	5195.5	0.0036

Figure 6 illustrates how A_B reacts to changing α and N, and it shows that the curves have a pattern that are mirror image of Fig. 2. For example, at $\alpha = 0.01$ (an extreme condition), A_B of different N converges at $\simeq 0.90$. Note here that $G(r_m)$ is the same as N. On the other hand, as α increases A_B decreases at approximately the same rate across all N until α reaches another extreme condition of 0.99; then A_B shows sudden increases, yet still at about the same rate for all N. The results of Area_B and A_B are summarized in Table 3.

Based on Fig. 6, we can easily visualize the maximum, minimum, and the point where A_B is approximately 50% of all possible area (in this case, $N \cdot \log(r_m)$). In fact, that is how we determined the curves to demonstrate in Fig. 5. As shown in Fig. 5, these points are approximately at $\alpha = 0.01$, 0.90, and 0.20, respectively. At $\alpha = 0.20$ and $A_b \simeq 50\%$, the curve is near linear and cut through diagonally the rectangle whose sides are $G(r_m) = N$ and $\log(r_m)$. As α decreases from 0.20, two things happen. First, $G(1)$ increases; and second, the curve moves northwesterly and causes Area_B to increase. Most of the drastic slope changes take place at the *first* few points on the curve; then the curve return to its linear form, though with a flatter slope. The slope of the linear portion decreases as α decreases. When α increases, the curve moves toward the southeastern direction, the slope of the linear portion of the curve also decreases, and the sudden jumps occur at the *last* few points away from the origin.

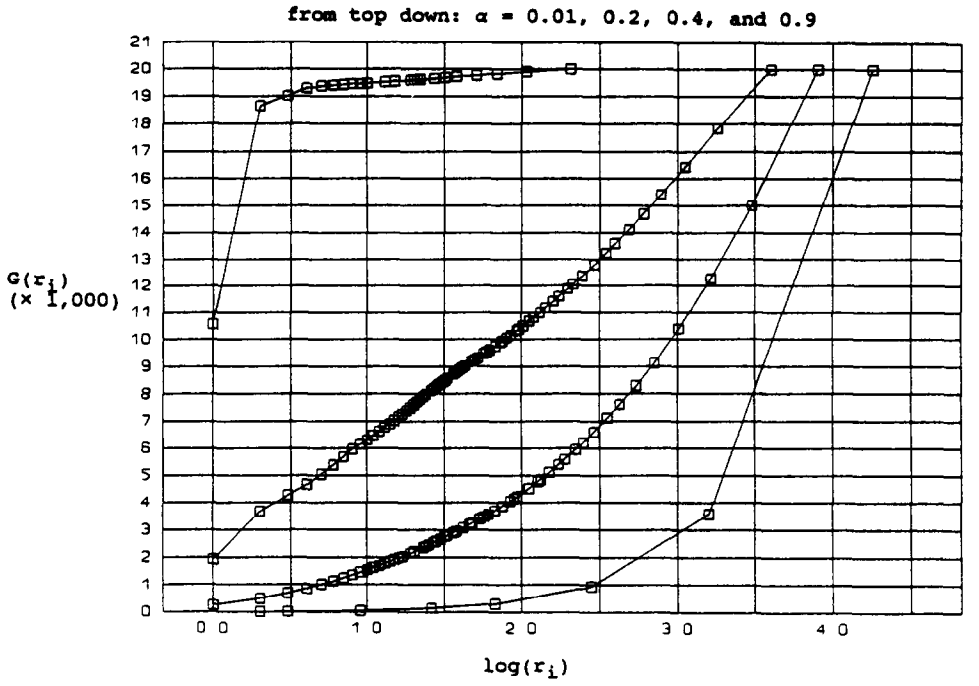

Fig. 5. Leimkuhler-Bradford's law with constant α ($N = 20{,}000$).

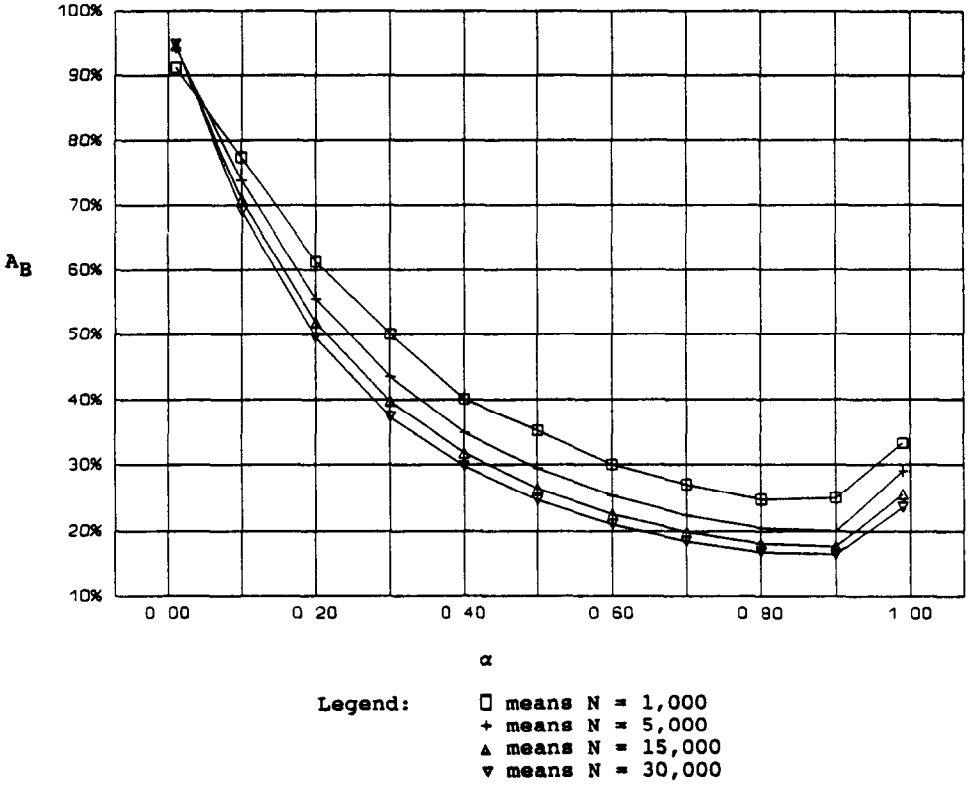

Fig. 6. Areas under Leimkuhler-Bradford's curve (A_B) with constant α.

Table 3. Simulation results of Bradford's law and Zipf's law with constant α

α	N (000)	log(r)	log($g(r)$)	Area$_B$	Area$_Z$	A_B	A_Z
0 01	1	0.954	2.728	869.9	1.441	0.9118	0.5537
0.10	1	1.944	2.569	1505.0	2.328	0.7742	0.4661
0.20	1	2.262	2.243	1387.5	2.559	0.6134	0 5044
0.30	1	2.480	1.959	1241.8	2.505	0.5007	0.5155
0.40	1	2.603	1.643	1044.3	2.361	0.4012	0.5521
0.50	1	2.695	1.544	951.0	2.182	0.3529	0.5245
0.60	1	2.786	1.204	835.9	1.894	0.3000	0.5646
0.70	1	2.855	1.114	770.6	1.669	0.2699	0.5249
0.80	1	2.903	0.903	719.1	1.163	0.2477	0 4435
0.90	1	2.957	0.699	739.5	0.947	0.2501	0.4581
0.99	1	2.994	0.477	996 6	0.708	0.3329	0.4958
0.01	5	1.643	3.419	7729.3	2.316	0.9409	0 4123
0 10	5	2.676	3.209	9883.8	3 827	0.7387	0.4457
0.20	5	2.996	2.808	8320.8	4 064	0.5555	0.4831
0.30	5	3.183	2.481	6935.9	3.932	0.4358	0.4979
0.40	5	3.304	2.037	5766.2	3.677	0.3490	0.5464
0.50	5	3.396	1.826	5001.8	3.348	0.2946	0.5399
0.60	5	3.471	1.415	4394.3	2.937	0.2532	0.5980
0.70	5	3.538	1.279	3960.6	2.551	0 2239	0.5638
0.80	5	3.597	1.079	3682.7	2.128	0 2048	0.5482
0.90	5	3.652	0.845	3661.6	1.674	0.2005	0.5424
0.99	5	3.695	0.477	5352 6	0.824	0 2897	0 4674
0.01	10	2.021	3.719	19072.6	2.934	0 9437	0.3904
0.10	10	2.999	3.479	21552.7	4.604	0.7187	0.4413
0.20	10	3.305	3.049	17572.8	4.827	0.5317	0.4790
0 30	10	3.482	2.704	14307 4	4.654	0.4109	0.4943
0.40	10	3.692	2.241	11826.1	4.343	0.3203	0.5249
0.50	10	3.692	1.973	10172 9	3.946	0.2755	0.5417
0.60	10	3.773	1.556	8855.8	3.425	0.2347	0.5834
0.70	10	3.841	1.398	7961.5	2 964	0.2073	0 5520
0.80	10	3.900	1.079	7370.7	2.420	0.1890	0.5751
0.90	10	3.955	0.845	7325.3	1.905	0.1852	0 5700
0.99	10	3.996	0.477	10770.8	0.938	0 2695	0.4921
0.01	15	2.173	3.899	30812.9	3.292	0 9453	0 3885
0.10	15	3.177	3.642	33737.3	5.096	0.7079	0.4404
0.20	15	3.483	3.189	27084.5	5.310	0.5184	0.4781
0 30	15	3.657	2.812	21828.4	5.113	0.3979	0.4972
0.40	15	3.773	2.342	17970.1	4.763	0 3175	0.5390
0.50	15	3.896	2.076	15322.2	4.306	0 2640	0.5361
0.60	15	3.950	1.663	13320.8	3.746	0.2248	0 5703
0.70	15	4.020	1.447	11983.8	3 243	0.1987	0.5575
0.80	15	4.078	1.176	11097 0	2.679	0.1814	0.5586
0.90	15	4.130	0.903	10996.1	2.092	0 1775	0.5609
0.99	15	4.172	0.477	16015.0	1.013	0.2559	0.5090
0.01	20	2.312	4.024	43742.4	3.552	0.9460	0.3818
0.10	20	3.295	3.758	46252.9	5.465	0.7019	0.4414
0.20	20	3.601	3.287	36674.3	5.672	0 5092	0.4792
0.30	20	3.780	2.894	29274.2	5.443	0 3872	0.4975
0.40	20	3.904	2.415	24039.3	5.051	0.3079	0.5357
0.50	20	3.996	2.137	20513.4	4.569	0.2567	0 5351
0.60	20	4.076	1.681	17826.6	3.966	0.2187	0 5789
0.70	20	4.145	1.491	16004.0	3.418	0.1931	0 5531
0.80	20	4.202	1.176	14797.3	2.837	0 1761	0.5742
0.90	20	4.256	0.903	14669.8	1.937	0 1723	0.5039
0.99	20	4.297	0.477	21098.5	1.000	0 2455	0.4877
0.01	25	2.391	4.121	56597 4	3.768	0 9468	0.3824
0.10	25	3.386	3.843	58938.7	5.755	0.6963	0 4423
0.20	25	3.698	3.368	46442.2	5.959	0.5023	0 4784
0.30	25	3.877	2.962	36876.2	5.708	0.3805	0 4970
0.40	25	4.000	2.470	30247.3	5.300	0.3025	0.5365
0.50	25	4.094	2.170	25685.6	4.781	0 2510	0.5381
0.60	25	4.173	1.699	22281.4	4.138	0 2136	0.5836
0 70	25	4.242	1.505	19976.8	3.550	0.1884	0.5561
0.80	25	4.300	1.204	18483.7	2.922	0 1719	0.5644
0.90	25	4.352	0.903	18321.9	2.009	0.1684	0 5113
0.99	25	0.394	0.477	26372.8	0.989	0.2401	0.4717

continued

Table 3. continued

α	N (000)	log(r)	log($g(r)$)	Area$_B$	Area$_Z$	A_B	A_Z
0.01	30	2.483	4.197	70564.8	3.942	0.9473	0.3783
0.10	30	3.465	3.911	71836.7	6.004	0.6911	0.4431
0.20	30	3.777	3.428	56089.2	6.201	0.4950	0.4789
0.30	30	3.956	3.013	44354.2	5.928	0.3737	0.4973
0.40	30	4.079	2.515	36381.9	5.499	0.2973	0.5360
0.50	30	4.173	2.207	30854.8	4.955	0.2465	0.5380
0.60	30	4.253	1.708	26746.7	4.285	0.2096	0.5899
0.70	30	4.322	1.505	23961.0	3.660	0.1848	0.5627
0.80	30	4.380	1.204	22163.9	3.013	0.1687	0.5713
0.90	30	4.432	0.903	21980.6	2.065	0.1653	0.5159
0.99	30	4.473	0.477	31608.6	0.951	0.2356	0.4459

5.2 *Decreasing entry rate*

Figure 7 is the composite graph of these three Bradford's curves using a decreasing function $\alpha(R) = A/\ln(R)$, $R = 1,2,\ldots 20{,}000$. The three curves are generated using $A = 1$, 1.25, and 2, with $A = 1.25$ to be the most linear of the three.

Figure 8 shows that A_B decreases as A (thus α) increases, independent of N. In fact, we selected $N = 1{,}000$, 20,000, and 30,000 for illustration purposes, and the three curves basically overlap each other. Table 4 summarizes the values of Area$_B$ and A_B at different levels of A and N. Since the "all possible area" can also be expressed as $N\cdot\log(r_m)$, it automatically increases when the number of iteration, N, increases. However, the overlapping curves in Fig. 8 imply that Area$_B$ (the nominal area) changes proportional to N; thus, A_B appears to remain unaffected by changing N. A simple regression analysis using $N = 20{,}000$ yields $A_B = 0.7244 - 0.1776A$, with $R^2 = 0.9843$. Similar to Fig. 5, the 50% curve ($A = 1.25$) in Fig. 7 is the most linear one, the minimal curve ($A = 2.00$) bends southeasterly, and the maximal curve bends northwesterly. Again, when α increases, both Area$_B$ and A_B decrease, and vice versa.

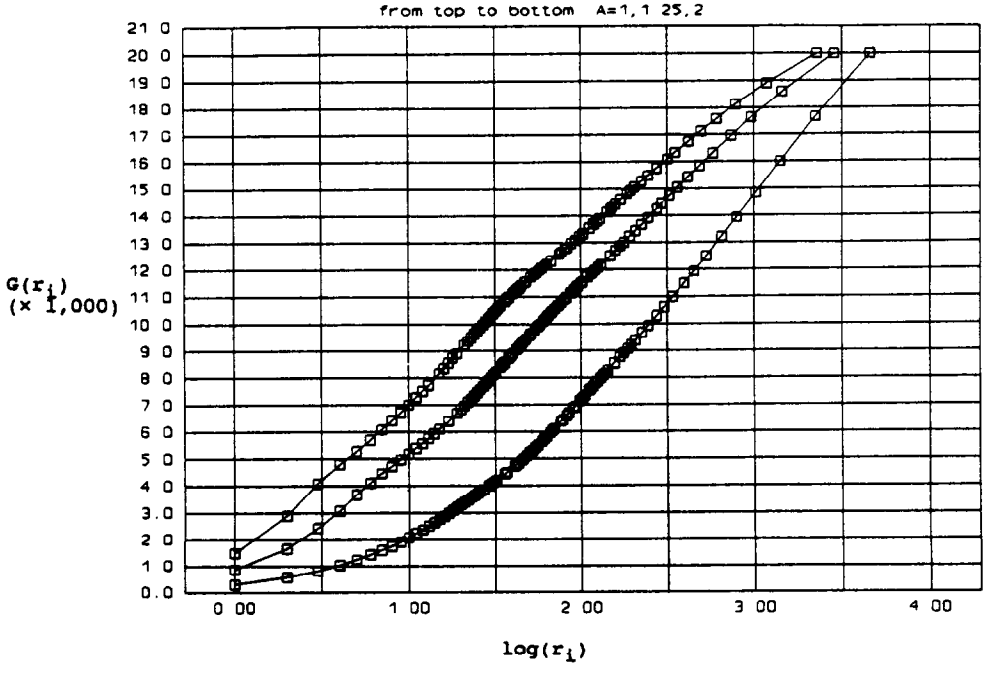

Fig. 7. Leimkuhler-Bradford's law with $\alpha(R) = A/\ln(R)$, $R = 1,2,\ldots,20{,}000$.

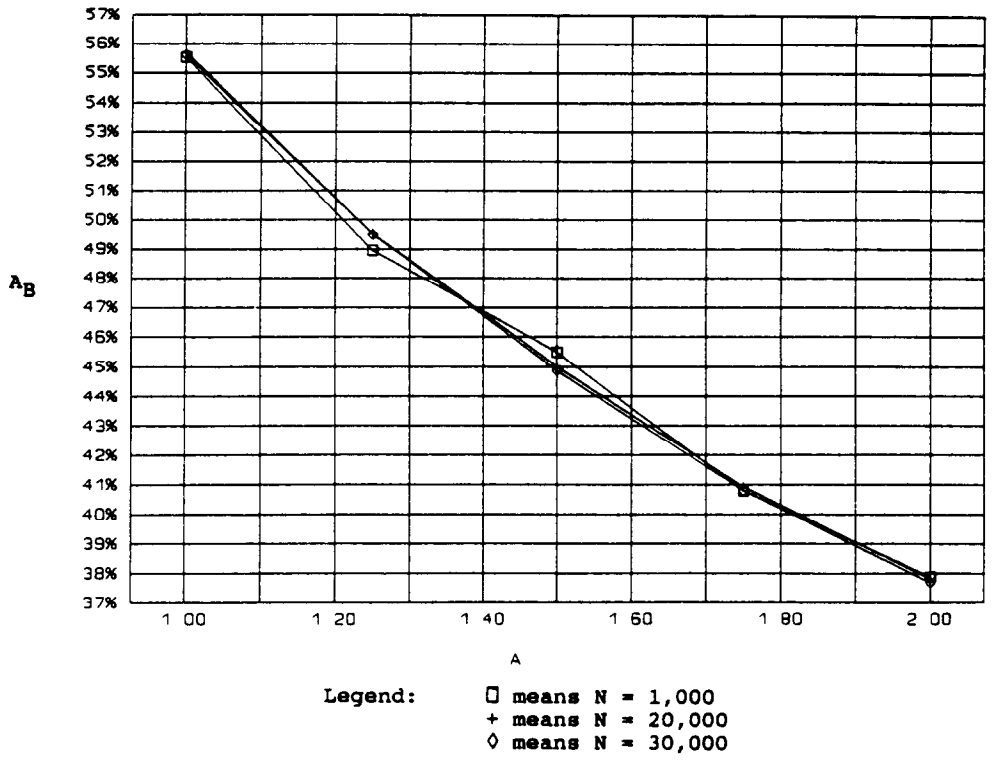

Fig. 8. Areas under Leimkuhler-Bradford's curve (A_B) with $\alpha(R) = A/\ln(R)$, $R = 1,2, \ldots, 20{,}000$.

Based on Fig. 8, the maximum, minimum, and the 50% points are determined to be at approximately $A = 1.0$, 2.0, and 1.25, respectively—the three curves we chose to include in Fig. 7.

6. COMPUTATIONAL EXPERIMENTATION OF ZIPF'S LAW

The Area under Zipf's curve (Area_Z) is calculated as follows:

$$\text{Area}_Z = \tfrac{1}{2}[(\log g(r_2) + \log g(r_1))(\log r_2 - \log r_1) + (\log g(r_3) + \log g(r_2))(\log r_3 - \log r_2) + \ldots + (\log g(r_m) + \log g(r_{m-1}))(\log r_m - \log r_{m-1})]. \quad (9)$$

We also define $A_Z = \text{Area}_Z/[(\log(g(r_m))(\log(r_m))]$, where the denominator is the largest possible Area_Z.

6.1 *Constant entry rate*

Several Zipf's curves are plotted in the same graph in Fig. 9. In addition to $\alpha = 0.01$, 0.30, 0.60, we add to our graph the other extreme point $\alpha = 0.99$.

Figure 10 shows how A_Z varies under different α and N. The pattern here is less clearly defined, especially when $N = 1{,}000$. However, if we flip the graph both horizontally and vertically, its general pattern is similar to that of Fig. 2. The effect of N shows greater dispersion and is not as pronounced as that of α. Also, when compared to Fig. 2 and Fig. 5, the maximum point locates at the near-center of the α spectrum, rather than at the extreme points. The values of Area_Z and A_Z are summarized in Table 3. The nominal values of the area, Area_Z, follows the general pattern of A_Z: as α increases, it increases also; however, after reaching a certain maximum point it eventually decreases.

Table 4. Simulation results of Bradford's law and Zipf's law with $\alpha(R) = A/\ln(R)$, $R = 1,2,\ldots,N$

A	N(000)	$\log(r)$	$\log(g(r))$	Area_B	Area_Z	A_B	A_Z
1.00	1	2.228	2.004	1237.2	2.615	0.5553	0.5856
1.25	1	2.326	1.792	1138.4	2.575	0.4894	0.6177
1.50	1	2.433	1.708	1107.0	2.510	0.4550	0.6039
1.75	1	2.501	1.591	1020.4	2.406	0.4080	0.6046
2.00	1	2.542	1.447	963.0	2.323	0.3788	0.6317
1.00	5	2.836	2.629	7908.0	4.153	0.5577	0.5570
1.25	5	2.927	2.405	7233.0	4.119	0.4942	0.5851
1.50	5	3.014	2.288	6777.1	4.052	0.4497	0.5875
1.75	5	3.086	2.124	6304.0	3.939	0.4086	0.6010
2.00	5	3.136	1.987	5949.7	3.830	0.3794	0.6147
1.00	10	3.101	2.900	17267.9	4.943	0.5568	0.5497
1.25	10	3.198	2.678	15798.8	4.912	0.4940	0.5735
1.50	10	3.276	2.562	14699.3	4.840	0.4487	0.5767
1.75	10	3.349	2.401	13693.6	4.721	0.4089	0.5871
2.00	10	3.403	2.260	12921.3	4.600	0.3797	0.5981
1.00	15	3.255	3.058	27166.2	5.444	0.5564	0.5469
1.25	15	3.348	2.829	24847.5	5.415	0.4948	0.5717
1.50	15	3.431	2.712	23168.7	5.342	0.4502	0.5741
1.75	15	3.505	2.551	21513.1	5.213	0.4092	0.5830
2.00	15	3.556	2.394	20231.1	5.087	0.3793	0.5976
1.00	20	3.358	3.176	37419.1	5.816	0.5572	0.5453
1.25	20	3.457	2.944	34234.8	5.786	0.4952	0.5686
1.50	20	3.537	2.828	31856.0	5.712	0.4503	0.5710
1.75	20	3.609	2.663	29544.1	5.584	0.4093	0.5810
2.00	20	3.661	2.497	27680.7	5.448	0.3780	0.5960
1.00	25	3.439	3.268	47879.2	6.113	0.5569	0.5439
1.25	25	3.538	3.031	43844.0	6.085	0.4957	0.5674
1.50	25	3.621	2.903	40739.6	6.011	0.4500	0.5718
1.75	25	3.692	2.732	37749.6	5.878	0.4090	0.5828
2.00	25	3.747	2.559	35432.5	5.743	0.3782	0.5989
1.00	30	3.508	3.341	58540.8	6.364	0.5563	0.5430
1.25	30	3.607	3.100	53549.2	6.333	0.4949	0.5664
1.50	30	3.690	2.980	49694.1	6.257	0.4489	0.5690
1.75	30	3.761	2.807	46031.2	6.122	0.4080	0.5799
2.00	30	3.816	2.635	43142.1	5.979	0.3769	0.5947

Based on Fig. 10, the minimum, the maximum, and the 50% points of A_Z are the ones illustrated in Fig. 9, namely, $\alpha = 0.01, 0.60, 0.30$, respectively. Figure 9 can be analyzed from several angles. First, the general negative slope remains to be the characteristics of these Zipf's curves; however, the slope flattens with the increase of α. Second, the initial "kink" in the Zipf's curve remains, but becomes less pronounced as α increases. Third, $\log(g(r))$ decreases when α increases, but $\log(r)$ increases when α increases; thus, the "largest area possible," as we have used the term previously, changes its shape from vertical rectangles to horizontal rectangles. Note that similarly to the results obtained from previous sections, when $\alpha = 0.60$ and $A_Z \simeq 50\%$, the curve is near linear. However, the curve is also near linear when $\alpha = 0.99$, partly due to the many fewer observation points (m, the maximum index, is 3).

6.2 *Decreasing entry rate*

Simulation results are summarized in Table 4. Figure 11 depicts the pattern of A_Z with respect to A and N. Since the minimal Area_Z is greater than 50%, only the minimum and maximum points are selected to plot the Zipf's curves in Fig. 12.

When these decreasing functions are evaluated at $A = 1.0$ and 2.0 with $N = 20{,}000$, the minimum values of α are $\simeq 0.1$ and $\simeq 0.2$, respectively, at the end of the iteration. Thus, it is no surprise that these two curves lie somewhere around the curve of the con-

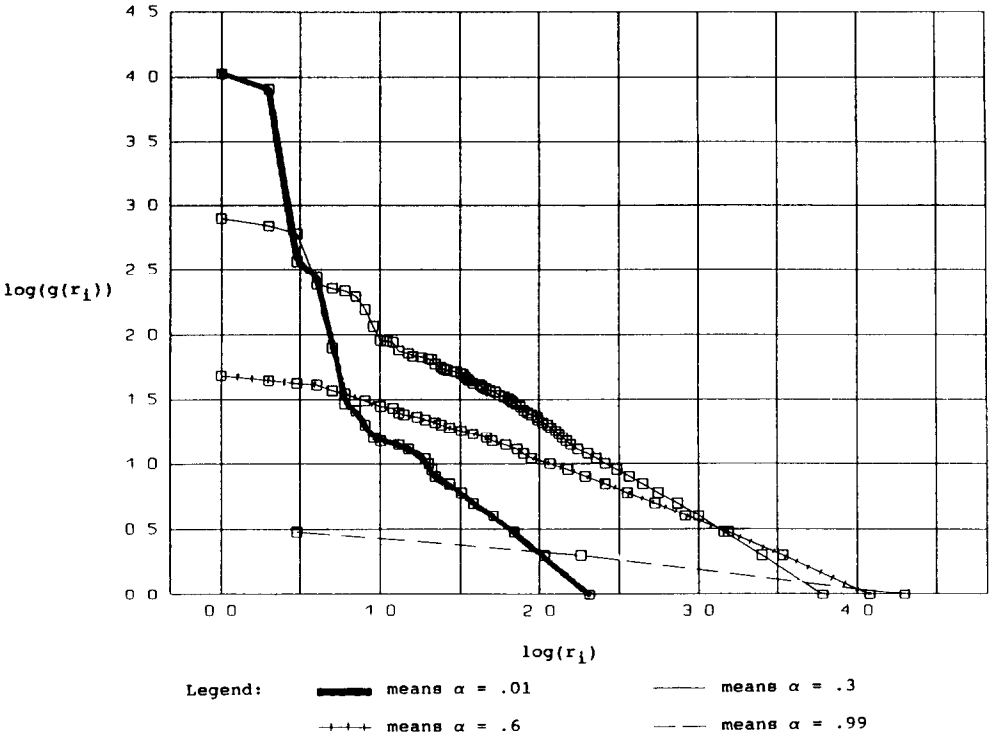

Fig. 9. Zipf's law with constant α ($N = 20{,}000$).

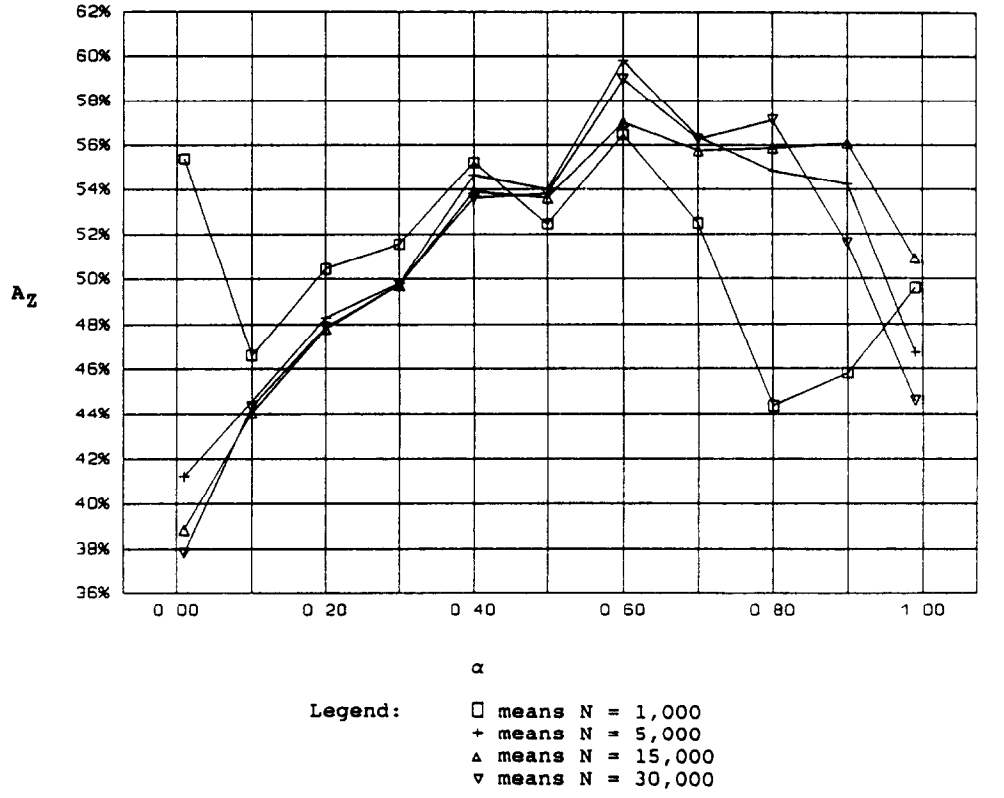

Fig. 10. Areas under Zipf's curve (A_Z) with constant α

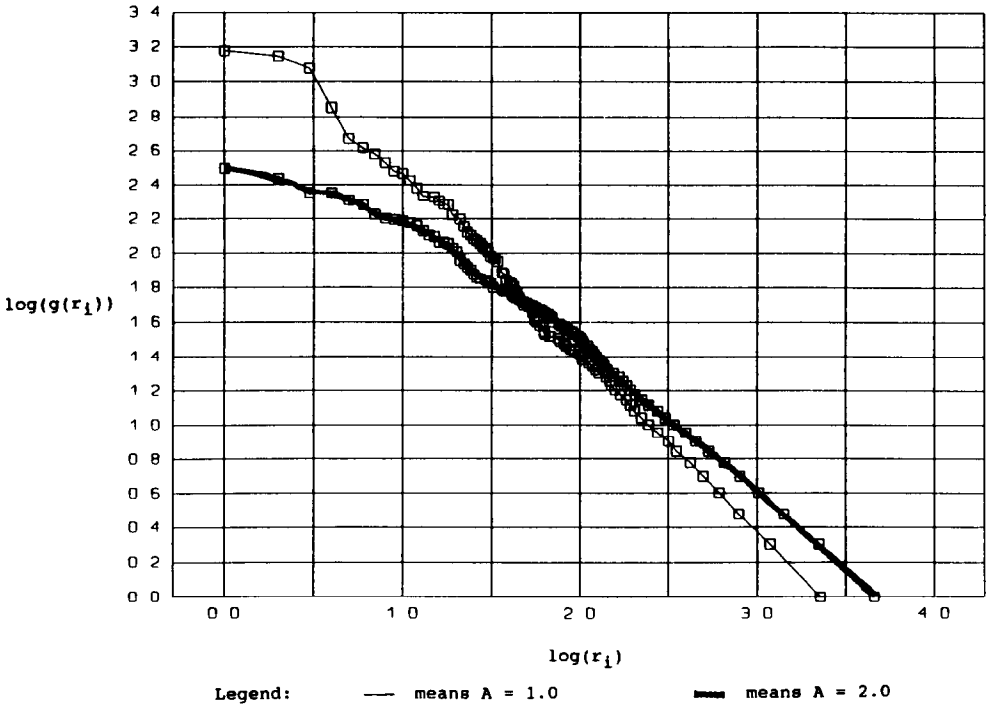

Fig. 11. Zipf's law with $\alpha(R) = A/\ln(R)$, $R = 1,2,\ldots,20{,}000$.

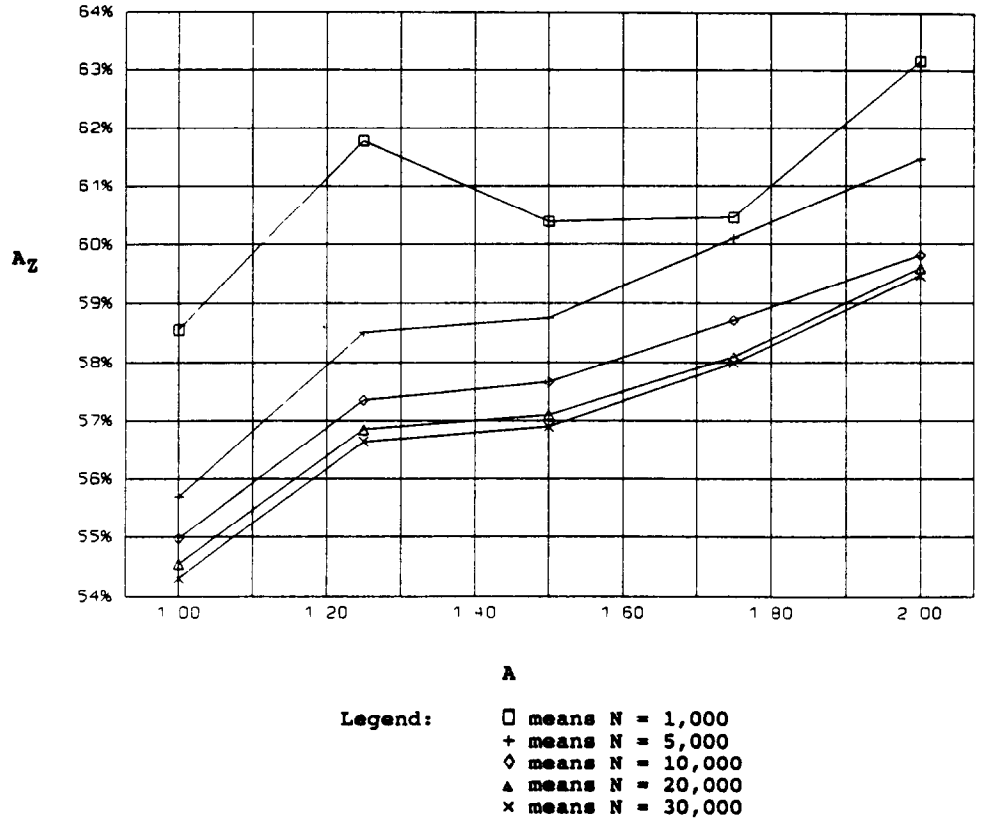

Fig. 12. Areas under Zipf's curve (A_Z) with $\alpha(R) = A/\ln(R)$ where $R = 1,2,\ldots,20{,}000$.

stant $\alpha = 0.30$ in Fig. 10. Most of the observations made in Section 6.1 still hold true—the flattened slope, the reduction of the initial "kink," the decreasing $\log(g(r))$, and increasing $\log(r)$—as α increases. Table 4 is the summary of values of Area_Z and A_Z under different A and N. Note that Area_Z and A_Z are inversely related.

7. SIGNIFICANT FINDINGS

7.1 *The effect of α and N*

As discussed in the previous sections, it is clear that the shapes of the curves in all three distributions are affected the most by the new entry rate α. All simulation results show that lower α raises the concentration level. This is because lower α implies higher chance of using old sources, which results in a higher concentration level. In the case of constant α, the number of iteration makes little difference in terms of the level of concentration. In the cases of decreasing function, larger N also causes the α to be smaller eventually, and the distributions are affected accordingly.

As observed in Section 6.2, decreasing $\alpha(R)$ can generate curves similar to those of constant α, depending on the minimum α in $\alpha(R)$. On the other hand, we did some preliminary experiments in mixing several levels of α within one simulation. For instance, we began with $\alpha = 0.01$ for 5,000 iterations, and then change α to 0.90 for another, 5,000 iterations. The results were compared with the ones begun with $\alpha = 0.90$ and then followed by $\alpha = 0.01$. The usage patterns are different, but patterns are difficult to detect in our limited tests. The effect of mixed α still awaits future research. On the other hand, α also affects other parameters in significant ways.

7.2 *Index number m*

Figure 13 indicates that regardless of the level of N, the simulation generates the maximum number of indexes (ranks) with constant $\alpha = 0.20$. Within the scope of our simulation, all indexes are reduced to be three when $\alpha = 0.99$, independent of the level of N. The maximum number of ranks m increases between $\alpha = 0.01$ to 0.20, then declines at an *increasing* rate as α continues to increase. On the other hand, Fig. 14 shows that n_m, the frequency of usage for the most used source (usually only one at this rank), declines continuously at a *decreasing* rate as α increases. This looks logical, since higher α implies lower chance of using old sources, which results in lower n_m.

7.3 *Average items per source*

Table 5 summarizes the average items per source (μ). Figures 15 and 16 show the relationships between μ and α. Apparently μ decreases as α increases, regardless of the level of N, and whether α is constant or a decreasing function. The difference is that in the case of constant α, μs of different N are very similar, whereas the separation is much greater in decreasing function. In either case, given that N is held constant in each simulation, higher α means lower concentration (due to increasing number of distinct sources used), which results in lower μ.

7.4 *Interpretation of the results*

At this point we need to revisit Simon's first assumption in his Model I. Since α is the rate of new entry, then given total items N, the total possible distinct sources $F(1)$ would be αN; thus, $\mu = 1/\alpha$, which is independent of N. Since lower α decreases $F(1)$, that means the same number of items must be distributed among a smaller set of sources; hence, higher concentration.

Simon's first assumption affects primarily $f(1)$ because α determines the probability of transferring numbers from $f(0)$ to $f(1)$. On the other hand, the second assumption determines the allocation of items among the previously used sources. Since increasing usage of a source increases its probability of being used again, with each iteration bypassing the first assumption, the most used sources have the probability of usage increasing exponentially. Given that a smaller α increases the usage of previously used sources, it is logical

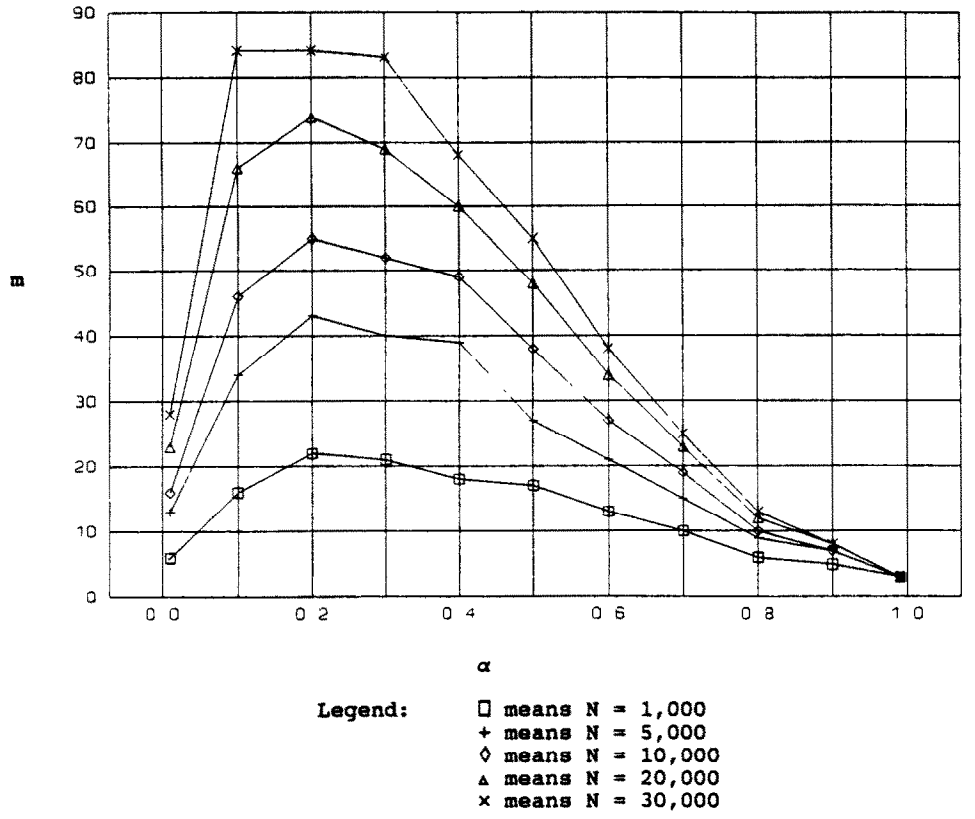

Fig. 13. Relationship between m and α.

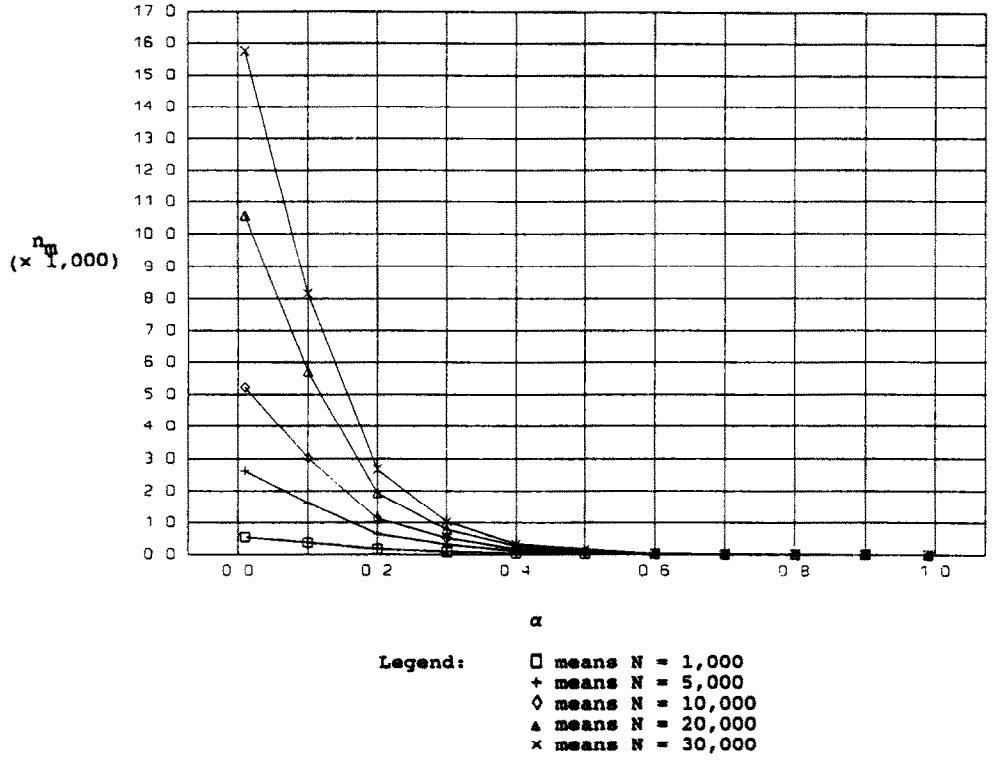

Fig. 14. Relationship between n_m and α.

Table 5a. Average items per source: Constant α

α	1K	5K	10K	15K	20K
0.10	11.3636	10.5485	10.0200	9.9734	10.1420
0.20	5.4645	5.0505	4.9554	4.9358	5.0125
0.30	3.3113	3.2830	3.2938	3.3054	3.3206
0.40	2.4938	2.4826	2.5094	2.5297	2.4969
0.50	2.0161	2.0105	2.0346	2.0305	2.0169
0 60	1.6367	1.6903	1.6855	1.6824	1.6810
0.70	1.3966	1.4493	1.4422	1.4310	1.4324
0.80	1.2516	1.2645	1.2596	1.2549	1.2550
0.90	1.1038	1.1131	1.1101	1.1119	1.1102

Table 5b. Average items per source: $\alpha(R) = A/\ln(R)$ $R = 1,2,.\quad ,N$

A	N = 1K	5K	10K	15K	20K
1.00	5.9172	7.2886	7.9239	8.3333	8.7719
1.25	4.7170	5.9172	6.3452	6.7355	6.9906
1.50	3.6900	4.8403	5.2938	5.5556	5.8038
1.75	3.1546	4.1017	4.4743	4.6904	4.9152
2.00	2.8736	3.6576	3.9494	4.1667	4.3611
Normalized by using A = 1.00 as base					
1.00	1.0000	1.0000	1.0000	1.0000	1.0000
1.25	0 7972	0.8118	0.8008	0.8083	0.7969
1.50	0 6236	0.6641	0.6681	0.6667	0.6616
1.75	0.5331	0.5628	0.5647	0.5628	0.5603
2.00	0.4856	0.5018	0.4984	0.5000	0.4972

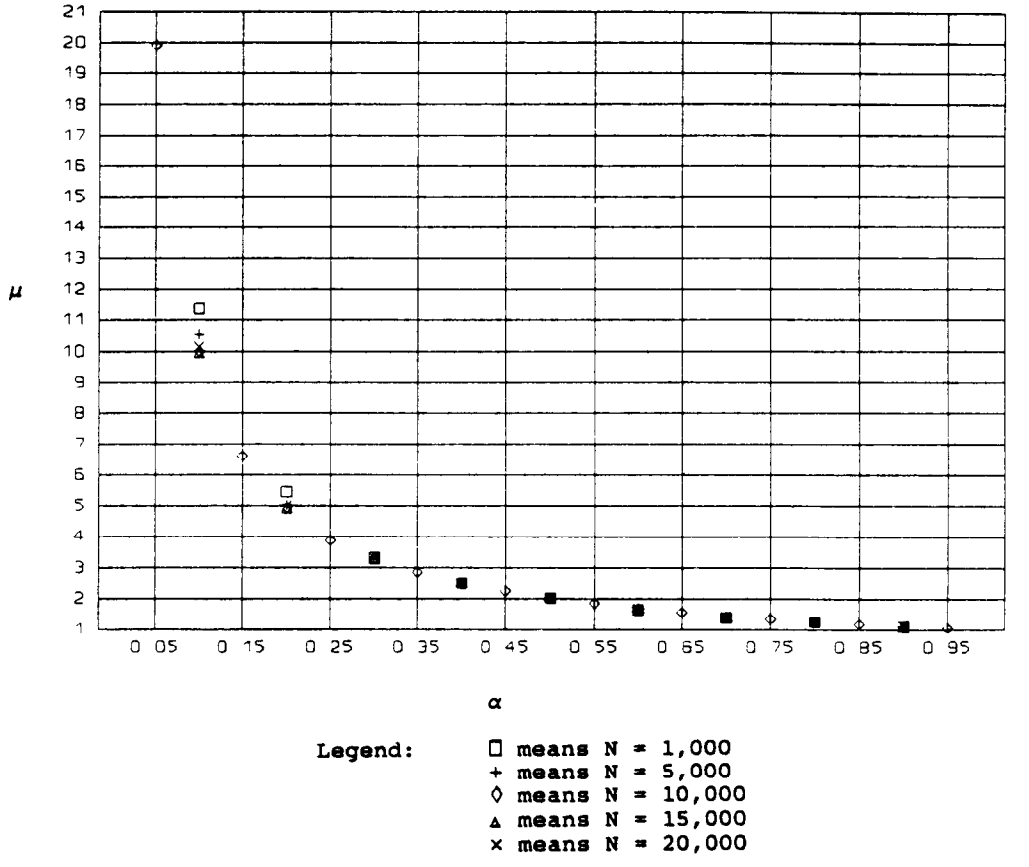

Fig. 15. Results: Average items per source (μ) with constant α.

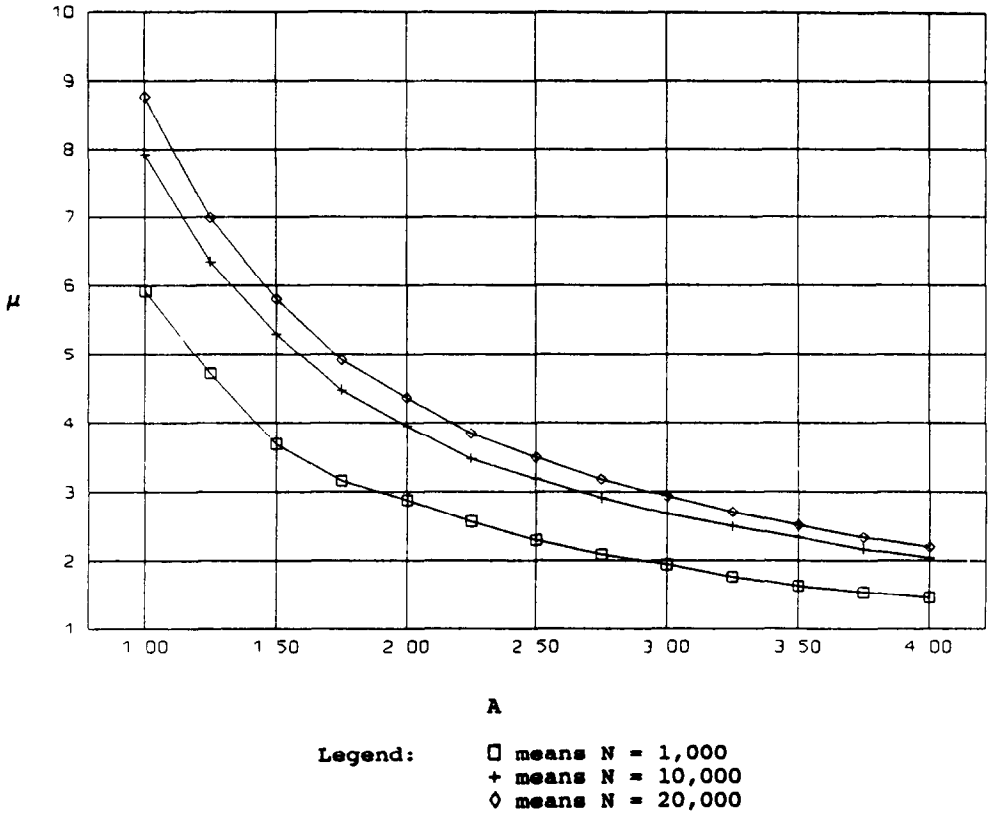

Fig. 16. Results: Average items per source (μ) with $\alpha(R) = A/\ln(R)$ where $R = 1,2,\ldots,20{,}000$.

that n_m should increase faster. This affects basically the region where $f(n_i) = 1$. Thus, changing α effectively changes the shape of the curves of these empirical laws.

It is not clear why m does not increase any faster with respect of N. Even at $N = 30{,}000$, m is less than 100, resulting in very large scattering patterns. It is also interesting to note that m is at or near maximum around $\alpha = 0.20$—an important point for all three empirical laws—albeit the reason for such a phenomenon is still under research.

8. CONCLUSIONS

In this paper we showed the importance of computational experimentation in deriving the behavior of the bibliometric distributions. Based on computational experiments, we demonstrated that the three bibliometric distributions can be simulated using Simon's generating algorithm. Furthermore, we analyzed the simulation results, and were able to conclude that the probability of new entry, α, was the most influential in determining the shapes of the curves and the characteristics of the distributions. On the other hand, the number of iterations, N, becomes a factor only when it affects α, as in a decreasing function $\alpha(R)$, $R = 1,2,\ldots,N$.

Acknowledgement—The authors thank the referees for their valuable comments on the early draft of the paper.

REFERENCES

Booth, A.D. (1967). A law of occurrences for words of low frequency. *Information and Control, 10*, 386–393.
Bradford, S.C. (1934). Sources of information on specific subjects. *Engineering, 137*, 85–86.
Chen, Y.S. (1989). Analysis of Lotka's law: The Simon-Yule approach. *Information Processing & Management, 25*(5), 527–544.

Chen, Y.S., & Leimkuhler, F.F. (1986). A relationship between Lotka's law, Bradford's law, and Zipf's law. *Journal of the American Society for Information Science*, *37*(5), 307–314.

Chen, Y.S., & Leimkuhler, F.F. (1987a). Bradford's law: An index approach. *Scientometrics*, *11*(3–4), 183–198.

Chen, Y.S., & Leimkuhler, F.F. (1987b). Analysis of Zipf's law: An index approach. *Information Processing & Management*, *23*(3), 171–182.

Chen, Y.S., & Leimkuhler, F.F. (1990). Booth's law of word frequency. *Journal of the American Society for Information Science*, *41*(5), 387–388.

Egghe, L., & Rousseau, R. (1989). *Proceedings of the Second International Conference on Bibliometrics, Scientometrics, and Informetrics*, U. of Western Ontario, London, Canada. Amsterdam: Elsevier.

Egghe, L., & Rousseau, R. (1990). *Introduction of informetrics*. Amsterdam: Elsevier.

Leimkuhler, F.F. (1967). The Bradford distribution. *Journal of Documentation*, *23*, 197–xxx.

Leimkuhler, F.F. (1988). On bibliometric modeling. *Informetrics*, 97–104.

Lotka, A.J. (1926). The frequency of distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–323.

Mandelbrot, B. (1953). An information theory of statistical structure of languages. *Proceedings of the Symposium on Applications of Communication Theory* (pp. 486–500). London: Butterworths.

Neuts, M.F. (1986a). An algorithmic probabilist's apology. In J. Gani (Ed.), *The craft of probabilistic modelling*. New York: Springer-Verlag.

Neuts, M.F. (1986b). Computer experimentation in applied probability. (Working paper 86–030, Systems and Industrial Engineering Department, University of Arizona).

Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika*, *42*, 425–440.

Simon, H.A. (1977). On judging the plausibility of theories. In Y. Ijiri & H.A. Simon (Eds.), *Skew distributions and the sizes of business firms* (pp. 109–134). Amsterdam: North-Holland.

Simon, H.A., & Van Wormer, T.A. (1963). Some Monte Carlo estimates of the Yule distribution. *Behavior Science*, *1*(8), 203–210.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.