Mathematical Modelling, Vol. 5, pp. 7–22, 1984
Printed in the USA.

0270–0255/84 $3.00 + .00

SIMULATING A NOMINATION PROCEDURE

M. KOCHEN, A. BLAIVAS, and R. BRUMBAUGH

University of Michigan
Ann Arbor, Michigan 48109

R. CRICKMAN

University of Minnesota
Minneapolis, MN 55455

Abstract—A snowball sampling procedure is simulated by computer. In snowball sampling, a small first-round sample of respondents, in this case editors of journals in several specialities, are asked to name second-round respondents to be contacted for a similar request for a third round, etc. In our survey, respondents were asked to nominate peer scientists for their contributions and expertise in their specialty. We used the simulation to estimate the effect of the number of rounds on the fraction of experts likely to be named, and to investigate the effect of other parameters. We found that it would take many rounds before every expert in a specialty is nominated. The simulation considered the effect of specialty subdivisions and showed how the distribution changes as the sample increases.

1. INTRODUCTION

A scientific community is intended to be more productive than a community whose members are producing independently. The structure of such communities plays a major role in their success. By structure, we mean the links among members of the community that are used for communication, coordination, propagation of influence, and control (Pool *et al.* [6]). In a scientific community, quality of research output is maintained by editors of scientific journals, who ask authors' peers to evaluate the manuscripts submitted for publication. Cole *et al.* [3] have shown that in peer evaluation of proposals, whom journal editors think of as a possible referee for a manuscript depends upon their personal knowledge of experts in the field of the paper. They can expand the pool from which to select referees by asking experts they know to recommend others. This constitutes a peer network among editors, potential authors, and referees. It partially characterizes the structure of a scientific community. It accounts in large measure for the effectiveness and efficiency of that community in adding to knowledge in its speciality.

To describe such networks, we selected several specialities and asked the editors of key journals to name the people in their field from whom they would most like to receive manuscripts or those whose judgment as referees they would greatly value. We then asked the people they named to submit their own nominations of experts in the field "whose work you try to keep up with, and whose competence, creativity, and judgment you respect."

The experts to whom the initial letters were sent out belong to seven different fields (information science, future studies, human systems management, general systems, polymer chemistry, differential geometry, and topology). Accordingly, all subsequent nominations

Key words: snowball sampling, nominating procedure, sociology of science, computer simulation, peer selection, social networks.

were separated into these fields. This procedure was repeated for four subsequent rounds. As a result, we obtained eight distributions of nominations: seven for seven different specialities and one aggregate which constituted the sum of all samples. We found that the distribution for some of the separate fields (differential geometry, topology, polymer chemistry, and general systems theory) fit well into a cumulative advantage distribution (CAD) which is often used to describe statistical processes with the so-called Matthew effect. The other fields and the overall distribution demonstrated a significant deviation from the CAD (Blaivas *et al.* [2]).

The general problem of sampling a network to describe its structure is of considerable interest and challenge. The technique of a chain letter of the kind described above (Kadushin [4]) appears to be useful, but presents a number of questions that must be answered before its validity can be assessed. Because of the considerable barriers to an approach via mathematical analysis, contrasted with the simplicity of simulating the process by computer, we decided to use the latter as a vehicle to answer some key questions.

2. METHODS

In order to formulate the key questions in language that the simulated procedure can be used to answer, we describe the simulation program first. We consider the following variables as input parameters to the program:

N = the number of individuals in the population being sampled. We name these N individuals $1, 2, 3, \ldots, N$, and denote the entire ordered set by P. The ordering is significant: we regard 1 and 3 to be the immediate "neighbors" of 2, and the left and right "neighbors" of 1 are N and 2, respectively. Thus, we consider our pool of nominees as a one-dimensional manifold, topologically similar to the ring (fig. 1).

M = the number of individuals selected by an individual as experts he or she values (20 in most cases).

RO = the number of rounds in the chaining procedure.

RA = range of nominations.

Each simulation run is specified by the following variables in addition to N, RA and M:

K = The number of specialized communities, each characterized by a triangular distribution on the circle of N points ($K = 2$ for the most part).

L_i = The location of the mode or apex for triangular distribution i, $i = 1, \ldots, K$; $1 < L_i < N$.

H_i = The height of the triangle at location L_i, $i = 1, \ldots, K$.

B_i = Half the base of the triangle. (Frequently, $W_i = RA$). Generally, the weights in the triangular distribution are positive integers. Points not covered by any triangular distribution are given weight 1.

Another input to the program is an *a priori* "distribution of visibility," which is reflected by the distribution of weights assigned to different individuals. At first, we started simulations with the simplest distribution, that is, weights equal to 1 for each individual in the pool; hence, if a randomly chosen nominator were allowed to nominate anyone in the pool, the probability of each being named would be exactly $1/N$.

The procedure for simulating the choice of nominees is as follows. A random sample of initial respondents is chosen, say #7 and #45. For each one, say for #7, a subpopulation of P is selected by including all individuals within RA to the left and RA to the right of the respondent. Thus, if $RA = 5$, then for #7 the subpopulation consisting of 2, 3, 4, 5, 6, 8, 10, 11, 12, 13 is sampled; the respondent him- or herself is excluded. A nominee is chosen at random from that subpopulation with no nominee chosen more than once by the same nominator until M individuals have been nominated.

For example, for respondent #7, the set of nominees that results may be $\{5, 9, 10\}$ for $M = 3$. This is now repeated for all other respondents. Suppose that, for #45, it is $\{38, 41, 43\}$, again for $M = 3$. If the respondents were closer together, say #15 instead of #45, then the nominees of #15 might be $\{10, 14, 16\}$ with #10 overlapping #7's nominees. Each individual may serve only once as respondent, regardless of how often he or she is nominated.

To simulate a simple population whose members enjoy unequal prominence, affecting the likelihood of their being nominated by others, we introduce non-uniform distributions, the simplest being a triangular distribution, as shown in Fig. 1. The reader can observe two triangular distributions close to each other, covering people with numbers 1–5 and 23–27.

Distribution of prominence or visibility among scientists has often been found to resemble a cumulative advantage distribution, perhaps explained with reference to a Matthew Effect (Merton [5]) or Law of Cumulative Advantage (Price [7]). The distribution of nominations in our actual sample has been compared to the theoretical cumulative advantage distribution in an earlier paper (Blaivas *et al.* [2]). To compare the results of our simulations, with and without weights representing unequal prominence, the program also calculated the cumulative advantage distribution for each simulation run, according to the formula,

$$f^*(n) = (m + 1)B(n, m + 2), \quad \text{where } B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)}$$

and $\Gamma(x)$ is the gamme function, $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$.

3. RESULTS

The questions we wanted to answer with this program were: 1) how many rounds are required before everybody in the population is named; and 2) how is the distribution of

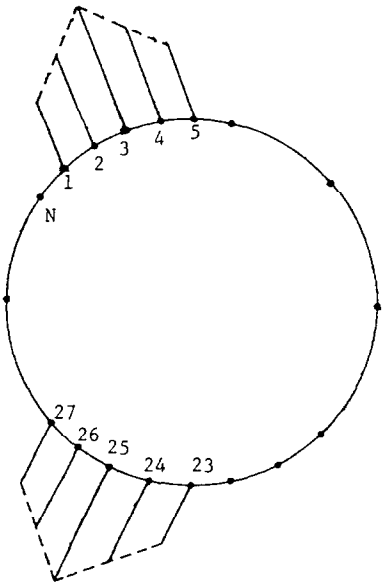

Fig. 1. Schematic representation of nomination pool as a ring with assigned weight of prominence to two groups of nominees in the form of two triangle distributions.

nominations affected by pool size, sample size, and the assumption that members of the pool are not all equally prominent?

A first question we explored with our model is the number of rounds required for everyone in the population to be named.

For a simulation in which $RA = +50$ and $M = 20$, and all weights are uniform, Fig. 2 shows the percentage of the population covered after each round for populations of 500 and 1000. The number of rounds required to approach complete coverage is large. After 12 rounds, 99.2% of the pool of 500 had been nominated, and 91.7% of the pool of 1,000 had been nominated.

Figure 3 shows the effect of introducing uneven weights for prominence into this population ($K = 2$, $L_1 = 11$, $L_2 = 40$, $H_1 = 20$, H, $B_1 = 10$, $B_2 = 10$) (see Methods). Here, after 12 rounds, 94.8% of the pool of 500 and 82.4% of the pool of 1000 had been nominated.

Both Figs. 2 and 3 show a steady levelling off in the curve as the sample approaches complete coverage of the population and new names become less frequent. For comparison, Fig. 4 shows results of our actual sampling procedure involving seven scientific fields. Since size of the population is unknown here, absolute numbers rather than percentages of the total are represented. The simulation does not duplicate all features of our actual study. Hence, a strict comparison is not possible. It is clear, however, that in the real sample as in the simulation, the proportion of new names to old names submitted in each round should decrease as the sample approaches complete coverage of the population. This effect can be seen in two of the fields studied: topology and differential geometry. Between our third and fourth rounds, the proportion of new to old names received dropped from 66.5% to 5% in topology and from 71.9% to 29.2% in differential geometry. We had earlier characterized these fields as tightly bounded, and these results suggest our sample is beginning to approach complete coverage in these fields. This effect does not appear in any of the other fields. There is a slight decrease for general systems theory, but increases of 11.7% for information science,

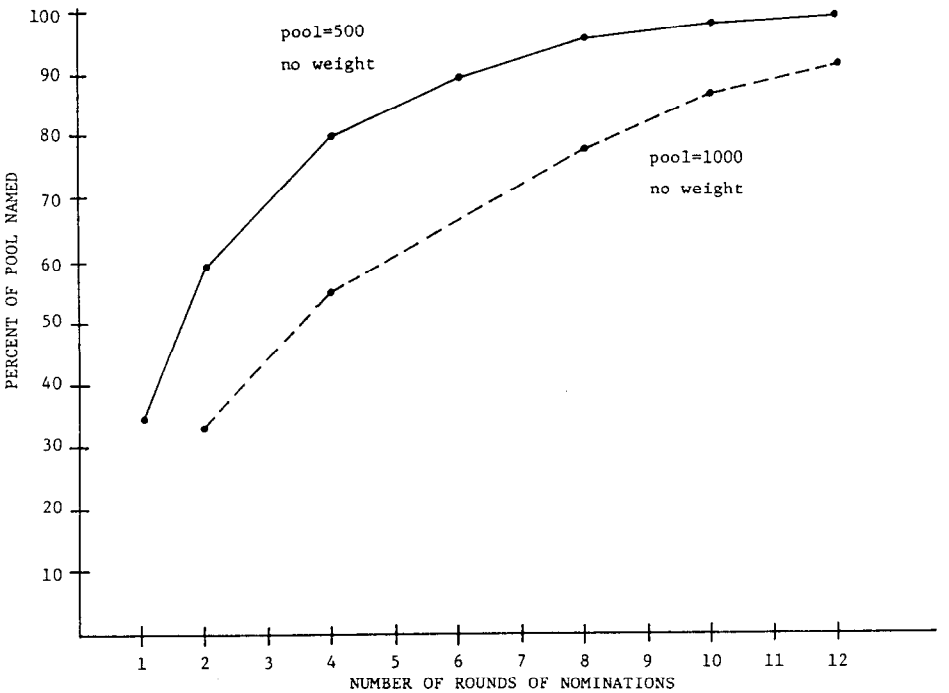

Fig. 2.

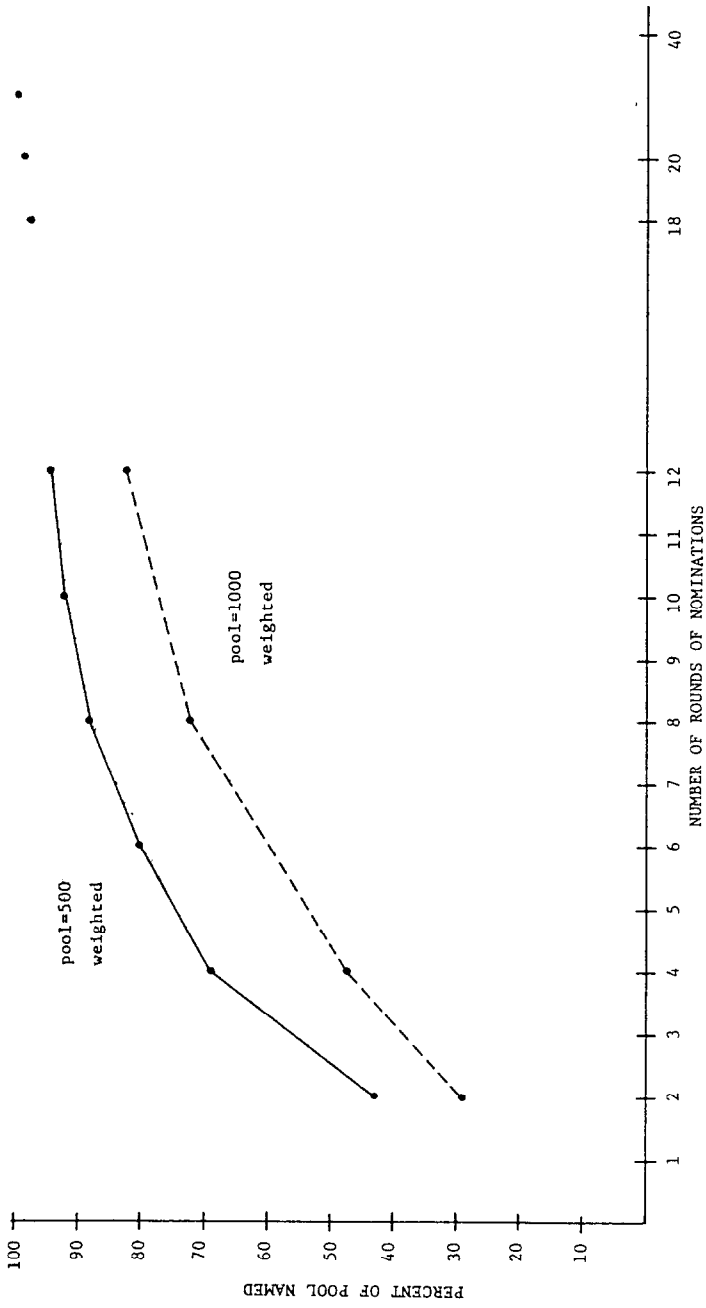
PERCENT OF POOL NAMED
100
90
80
70
60
50
40
30
20
10
1
2
3
4
5
6
7
8
9
10
11
12
18
20
40
NUMBER OF ROUNDS OF NOMINATIONS
pool=500
weighted
pool=1000
weighted

Fig. 3.

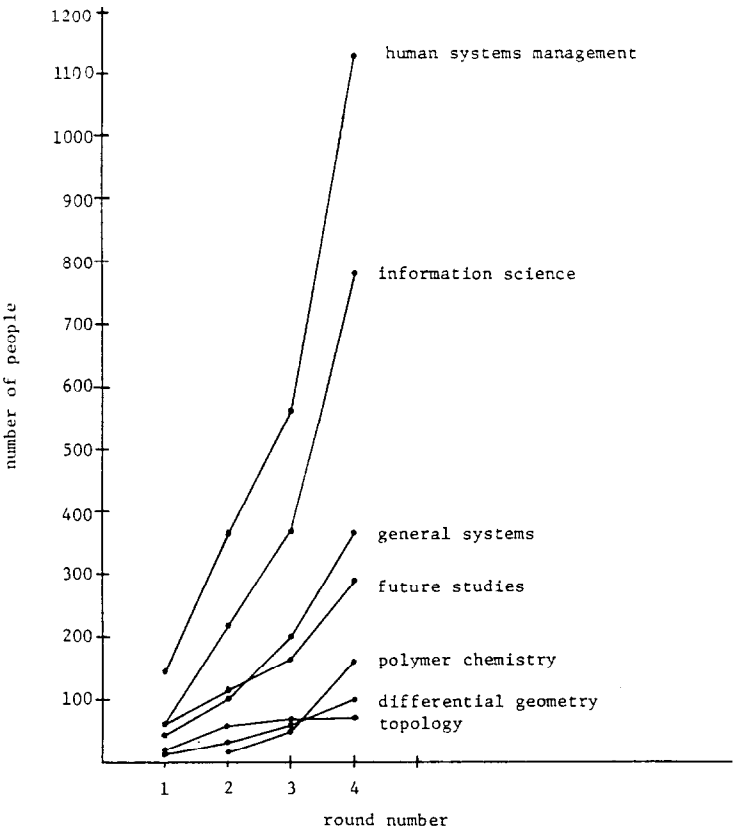

Fig. 4.

12.6% for human systems management, and 33.3% for future studies. An increase in the proportion of new names to old could be explained on the assumption that these are weakly bounded fields: early rounds might begin with a central group of workers among whom mutual recognition is higher, later rounds include more nominations by peripheral numbers each reaching into areas not closely related to, or likely to be named by, other nominators. This is consistent with our earlier characterization of these fields as diffuse and weakly bounded.

Distribution of nominations

In our study, some people by the end of four rounds had received many nominations while most had been named only once. The actual distribution of nominations in each speciality field is presented in Fig. 5.

As we expected, the distribution of nominations approximates, and for certain fields fits, the theoretically derived cumulative advantage distribution. These results have been discussed (Blaivas *et al.* [1]; Blaivas *et al.* [2]) and interpreted with reference to the "Matthew Effect" (Merton [5]), and discussions of cumulative advantage distribution by Price [7] and others.

For comparison, Fig. 6 presents the distribution of nominations obtained in our simulation after four rounds, for the population of 500 and 1000. This does not resemble either the cumulative advantage distribution or any actual distribution obtained in our study.

When weights signifying differences in prominence are added to the simulation (see Fig. 3), the distribution of nominations shown in Fig. 7 does approximate those we actually

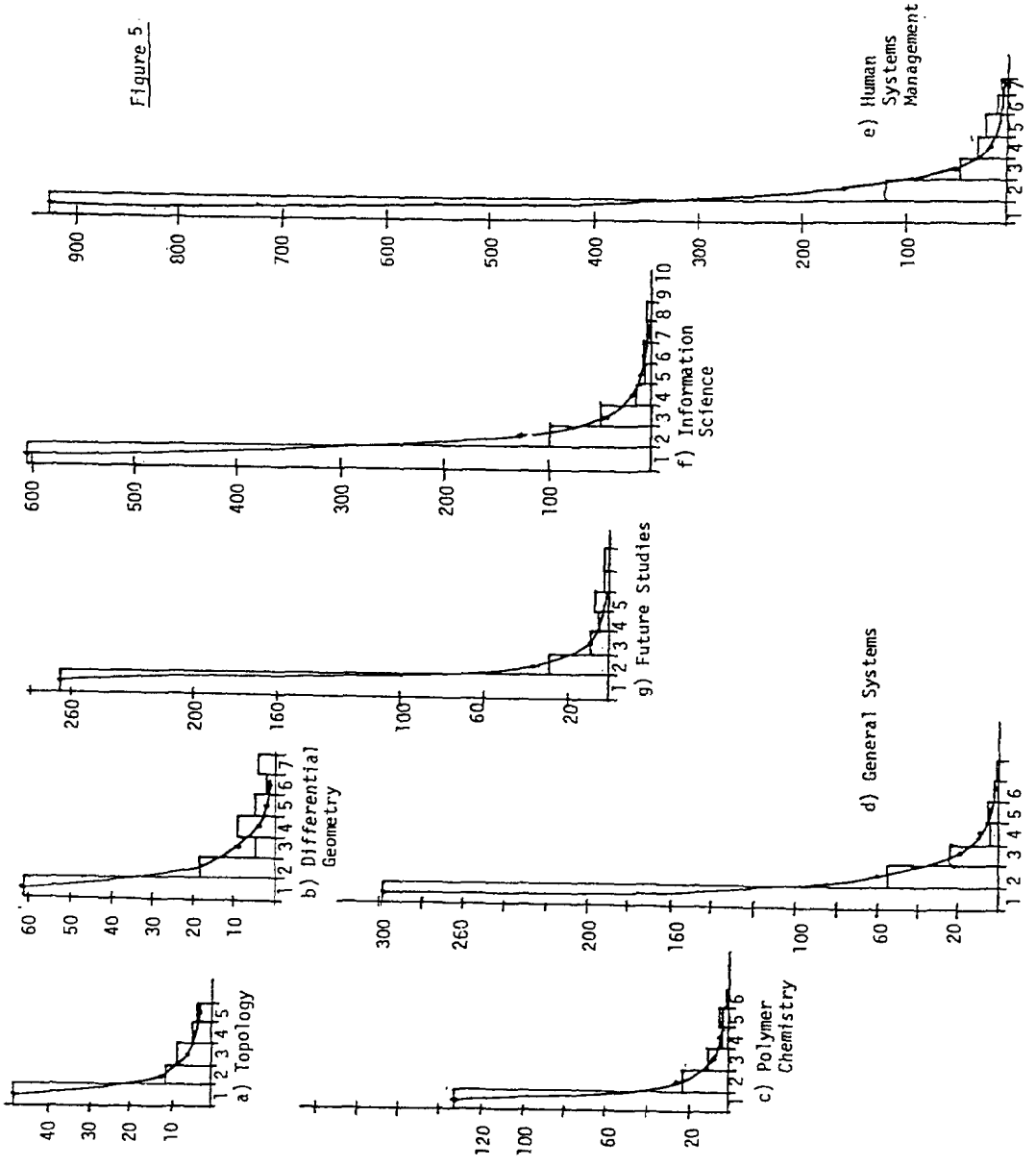
Figure 5
a) Topology
b) Differential Geometry
c) Polymer Chemistry
d) General Systems
e) Human Systems Management
f) Information Science
g) Future Studies

Fig. 5.

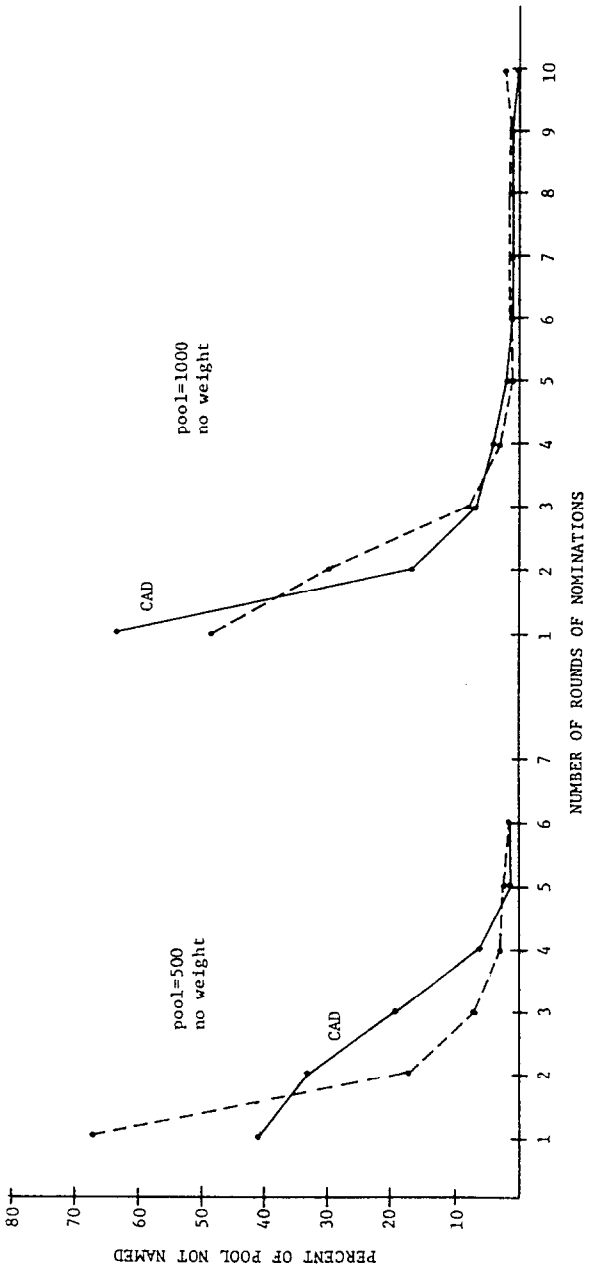
PERCENT OF POOL NOT NAMED
80
70
60
50
40
30
20
10
NUMBER OF ROUNDS OF NOMINATIONS
1
2
3
4
5
6
7
8
9
10
pool=500
no weight
CAD
pool=1000
no weight
CAD

Fig. 6.

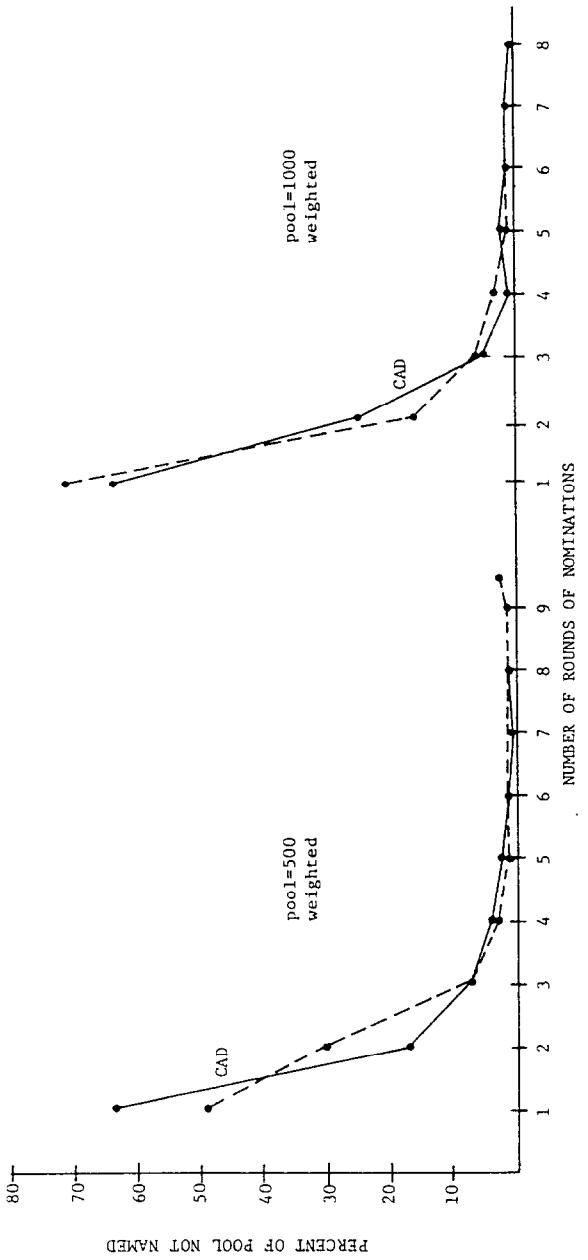
pool=1000
weighted
CAD
pool=500
weighted
CAD
NUMBER OF ROUNDS OF NOMINATIONS
PERCENT OF POOL NOT NAMED
80
70
60
50
40
30
20
10
1
2
3
4
5
6
7
8
9

Fig. 7.

obtained. By adjusting the weights, we have found that we can obtain a distribution to match the cumulative advantage distribution, or any of the actual distributions we obtained in Fig. 4.

Effect of pool size and number of rounds on distribution

Figure 8 shows the distribution of nominations after 2, 4, 8 and 12 rounds for populations of 500 and 1000, displaying the effect on distribution as the sample begins to approach full coverage of the population. As this happens, the peak of the distribution begins to shift. For example, by round 8 when 90% of the pool of 500 have been nominated, the number who have been named only once is smaller than the number with two nominations. This effect will of course be automatic as the sample approaches complete coverage of the population. It must be taken into account when results are interpreted (e.g., where distributions in two populations of very different size are compared).

While precise conclusions cannot be drawn from the simulation, it is clear that a similar shift in frequency distribution should appear as the sample approaches full coverage of the population: this is the other side to the change noted already, that the proportion of new to old names received in each round should decrease. Again it is topology and differential geometry which seem to show signs of approaching the limit of their populations, with 60% and 64% of the nominees receiving only one nomination, considerably lower than the other fields (see Fig. 5).

Effect of increasing the number of specialties in the pool

In reality, the same field may be partially subdivided into various speciality groups, individuals gaining prominence in one or another specialty. To approximate this situation, our simulation of differing degrees of prominence assumed two specialties as described under "simulation." Because specialty divisions or their equivalent are relevant to any actual study, we have examined the effect which the number of specialty divisions in the pool has on the distribution of nominations. It turns out that increasing the number of specialties does affect the frequency distribution, although not as dramatically as an increase in the number of rounds. An increase in the number of specialties primarily affects the first column of the distribution (see following tables) and decreases the number of persons having exactly one nomination, which is to be expected.

Tables 1, 2a–d give a dynamic representation of changes which occurred in the distribution when we introduced new specialties one by one. Each specialty was represented by a triangle distribution of weights: 1, 2, 3, 4, 6, 10, 8, 6, 4, 2, 1. Only starting positions varied for different specialties. The first specialty always started from place #1, the second from place #31.

Tables 2a, b, c and d represent the result of nominations with 2, 3, 4, and 5 specialties, respectively.

It is clearly seen that the introduction of new specialties produces a distribution further removed from the cumulative advantage.

Effect of the width of the specialties upon the distribution

Table 3a represents the results of a run under the same conditions as the experiment in Table 2d (with 5 specialties) with the distinction that widths and heights have been increased. Now they look like this:

1, 2, 4, 6, 8, 10, 12, 14, 16, 14, 12, 10, 8, 6, 4, 2, 1.

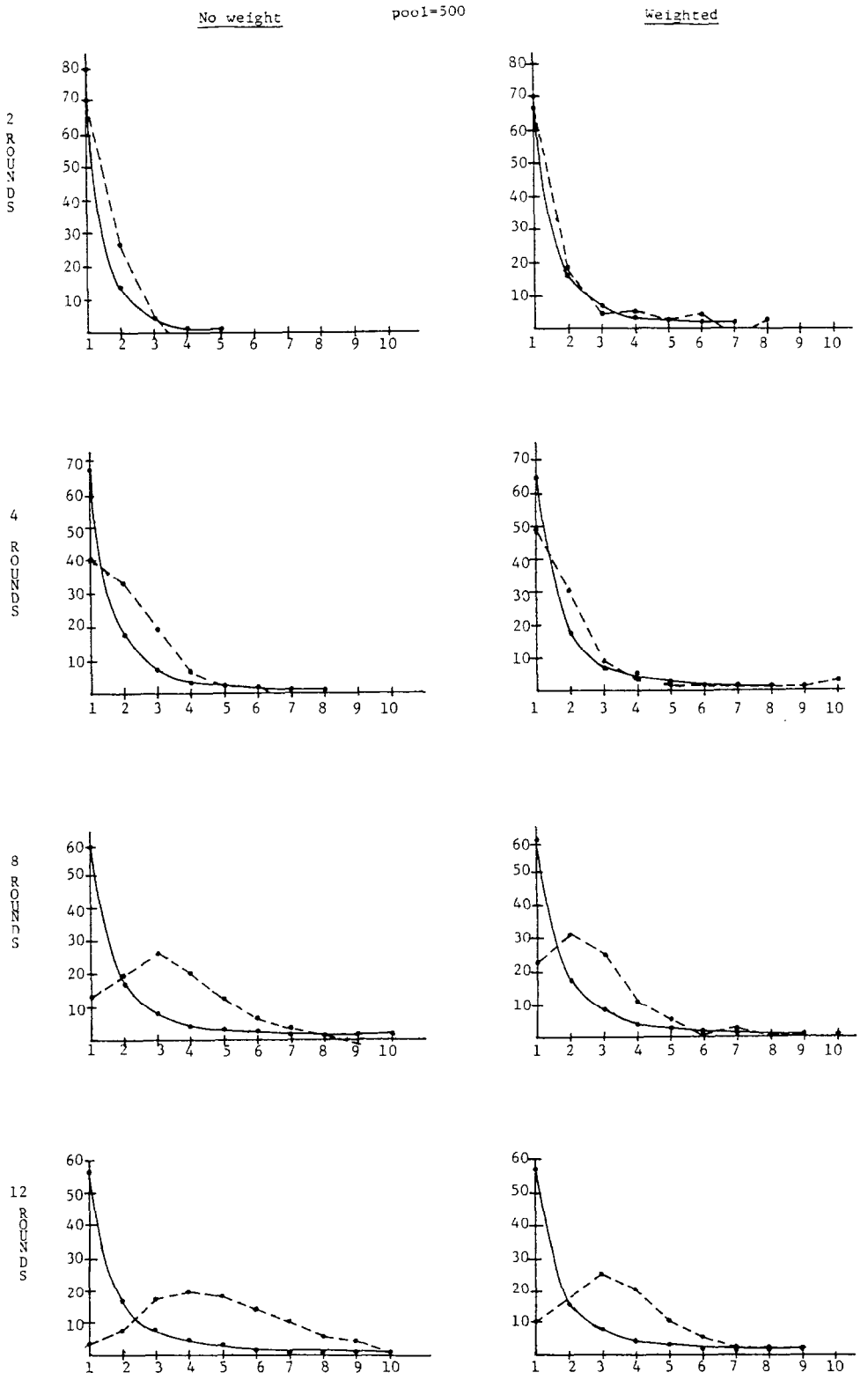
No weight
pool=500
Weighted
2 ROUNDS
4 ROUNDS
8 ROUNDS
12 ROUNDS

Fig. 8.

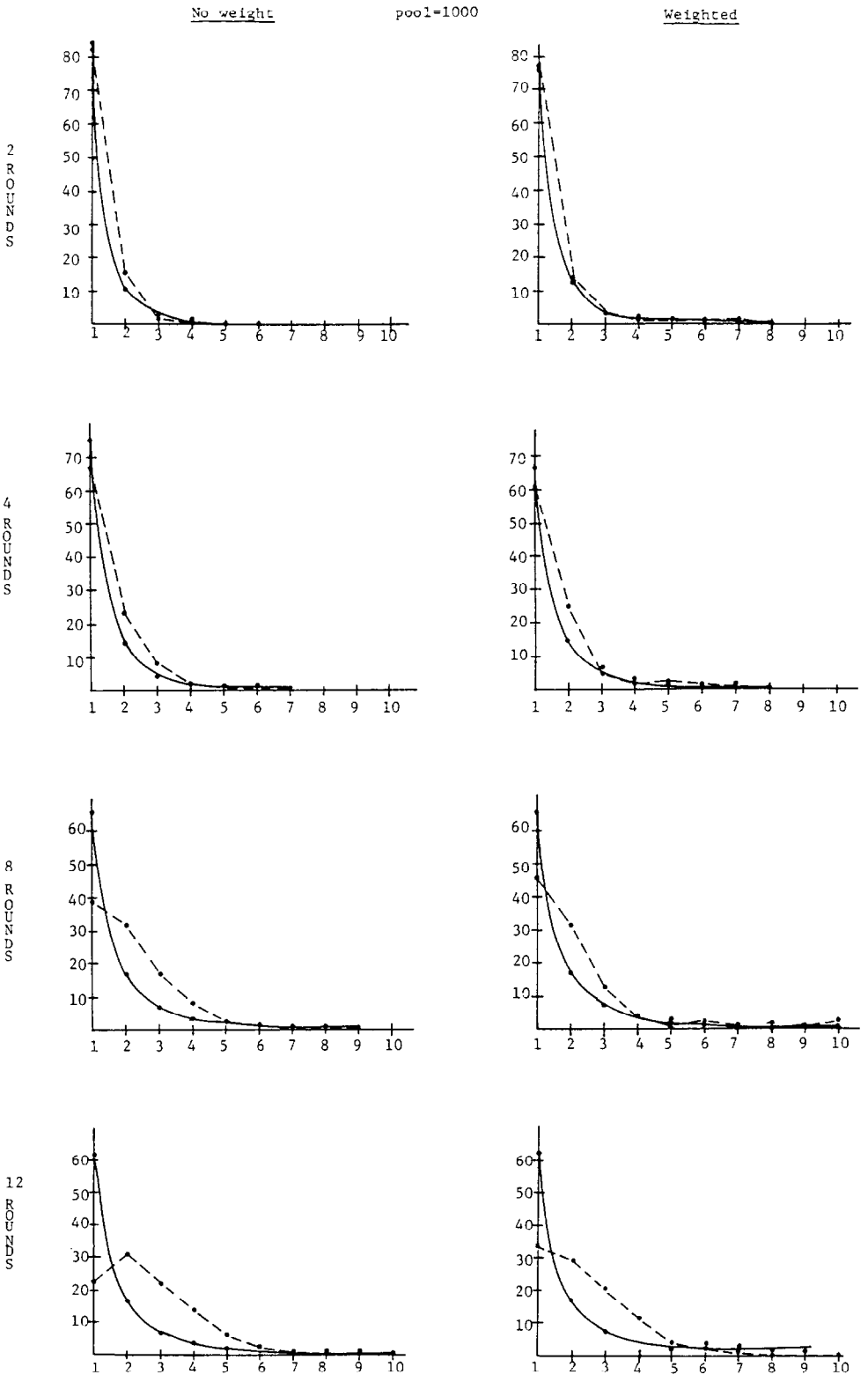
No weight
pool=1000
Weighted
2 ROUNDS
4 ROUNDS
8 ROUNDS
12 ROUNDS

Fig. 8. (cont.)

Thus, each specialty consists of 15 members with weights higher than others in the pool.

Table 3b gives this for 5 specialties with the following distribution of weights within each one:

1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 18, 16, 14, 12, 10, 8, 6, 4, 2, 1.

Comparing these three tables we can see that there is no dramatic change in distribution with variation of this parameter.

Table 1

No weight

Distribution:

210	77	12	0	0	0	

Number of observations = 299
Average = 1.34
Standard deviation = 0.30

Frequencies:

0.70	0.26	0.04	0.00	0.00	0.00	0.00

Cumulative advantage:

0.80	0.13	0.04	0.01	0.01	0.00	0.00

Table 2a

2 specialties
2 rounds, 500 people

Distribution:

176	66	16	2	1	2	0	0	1	1

Number of observations = 265
Average = 1.51
Standard deviation = 1.10

Frequencies:

0.66	0.25	0.06	0.01	0.00	0.01	0.00	0.00

Cumulative advantage:

0.75	0.15	0.05	0.02	0.01	0.01	0.00	0.00

Table 2b

3 specialties
2 rounds, 500 people

Distribution:

173	53	13	7	0	1	3	1	1	1

Number of observations = 253
Average = 1.58
Standard deviation = 1.62

Frequencies:

0.68	0.21	0.05	0.03	0.00	0.00	0.01	0.00

Cumulative advantage:

0.73	0.15	0.05	0.02	0.01	0.01	0.00	0.00

Table 2c

4 specialties
2 rounds, 500 people

Distribution:

162	60	13	7	3	2	1	1	1	0

Number of observations = 250
Average = 1.60
Standard deviation = 1.33

Frequencies:

0.65	0.24	0.05	0.03	0.01	0.01	0.00	0.00

Cumulative advantage:

0.73	0.16	0.06	0.02	0.01	0.01	0.00	0.00

Table 2d

5 specialties
2 rounds, 500 people

Distribution:

147	54	16	7	2	5	3	1	0	0

Number of observations = 235
Average = 1.70
Standard deviation = 1.61

Frequencies:

0.63	0.23	0.07	0.03	0.01	0.02	0.01	0.00

Cumulative advantage:

0.71	0.16	0.06	0.03	0.01	0.01	0.01	0.00

Table 3a

5 specialties
Width = 15

Distribution:

134	39	16	11	9	5	3	0

Number of observations = 217
Average = 1.84
Standard deviation = 1.96

Frequencies:

0.62	0.18	0.07	0.05	0.04	0.02	0.01	0.00

Cumulative advantage:

0.69	0.16	0.06	0.03	0.02	0.01	0.01	0.00

Separation of specialties

We have found that separating specialties in the pool significantly affects the distribution of nominations. For example, one can compare these two runs (I and II). Both were performed upon a pool of 500 nominees with 10 nominations allowed for each nominee. We had 4 specialties in each run, and each specialty was represented by a 1, 2, 4, 6, 10, 8, 6, 4, 2, 1 triangular distribution of weights. All other nominees outside of these 4 specialties had equal (= 1) weight assignments. In the first run, the starting points of the specialties were 1, 31, 51, 71; in the second, 1, 51, 101, 151. The difference is seen in Tables 4a and 4b.

Table 3b

5 specialties
Width = 20

Distribution:

127	34	16	10	12	4	3	

Number of observations = 207
Average = 1.91
Standard deviation = 2.24

Frequencies:

0.61	0.16	0.08	0.05	0.06	0.02	0.01	0.00

Cumulative advantage:

0.68	0.17	0.06	0.03	0.02	0.01	0.01	0.01

Table 4a

Run I
2 rounds, 500 people

Distribution:

148	65	18	5	4	1	2	100

Number of observations = 244
Average = 1.64
Standard deviation = 1.20

Frequencies:

0.61	0.27	0.07	0.02	0.02	0.00	0.01	0.00

Cumulative advantage:

0.72	0.16	0.06	0.03	0.01	0.01	0.01	0.00

Table 4b

Run II
2 rounds, 500 people

Distribution:

184	53	12	3	5	1	2	110

Number of observations = 262
Average = 1.53
Standard deviation = 1.32

Frequencies:

0.70	0.20	0.05	0.01	0.02	0.00	0.01	0.00

Cumulative advantage:

0.74	0.15	0.05	0.02	0.01	0.01	0.00	0.00

It is seen from the comparison that the second distribution more closely resembles the cumulative advantage distribution.

Dependence of the simulations upon the statistical procedure for generating random numbers

It is obvious that, since the nominations are "made" by the random number generator, we may expect some fluctuation in the resulting values with different choices of the initial seed or selection of another random number generator. A special experiment was conducted to ensure that the fluctuation range was small enough to be ignored.

For two simulations ($N = 500$, two weighted specialties, $R = 2$ and $R = 10$), we ran the same experiment 25 times. Each time the initial seed was selected from an independent table of random numbers created by a different random number generator. It turned out that fluctuations for the histogram values were small. For example, for $x = 0.67$ the standard deviation was 0.025. The variance did not grow with the increasing number of rounds.

4. CONCLUSION

The simulation procedure we explored in this publication has proved a useful tool for examining the important type of sampling procedure we used in our earlier study. Basically, the program showed that we can easily simulate the most fundamental properties of the nomination procedure: the relatively open character of choosing the nominee, and uneven degrees of prominence which lead to an observed distribution closely related to the cumulative advantage distribution. We discovered that we needed a surprisingly large number of rounds to have everyone nominated. Our simulation has considered the effect of specialty subdivisions in the population and showed how distribution begins to change as the sample extends to a larger proportion of the total population.

REFERENCES

1. A. Blaivas, R. Brumbaugh, R. Crickman, and M. Kochen, Consensuality of Peer Nominations Among Scientists. *Knowledge*, 1982.
2. A. Blaivas, R. Crickman and M. Kochen, Distribution of Scientific Experts as Recognized by Peer Consensus, Scheduled for publication in *Scientometrics*, March 1982.
3. J. Cole, S. Cole, and G. A. Simon, Chance and Consensus in Peer Review, *Science* **214**, 4523, 881–886 (1981).
4. C. Kadushin, *The American Intellectual Elite*. Boston: Little and Brown (1974).
5. Robert K. Merton, *The Sociology of Science*. Chicago: University of Chicago (1973).
6. I. De Sola Pool and M. Kochen, Contacts and Influence. *Social Networks* **1**, 5–51 (1978).
7. D. J. De Solla Price, A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American Society for Information Science* (1976).