Who's afraid of research assessment?

Gangan Prathap

Although the field of scientometrics now offers well-tested procedures for some measure of quantitative assessment of research performance, these are largely left unused in our country when we attempt exercises to assess the performance of individuals or institutions. This is baffling in a country that is so comfortable with its obsession with cricket and cricket statistics. The present analysis is based on data from the SCOPUS database, and this approach has the potential to offer interesting sociological insights into the scientific productivity of individuals, research institutes and research agencies.

The skewness of unfairness and scientific productivity

'Life is unfair', said John F. Kennedy famously. Human ability ranges widely and is distributed in a highly nonlinear fashion within a population. It has been known that patterns exist, which go beyond conventional Gaussian or normal distribution. Rank-order statistics based on power-law distributions is used to describe this. Scientometrics suggests that an area of intellectual activity that is most easily amenable to quantification is the production of research output as measured by publications in the open literature. The ecology of this enterprise, where a large number of scientists work, about half of them publish, but only a few account for the highly cited work, is complex. Slowly, soft laws have emerged, where the role of power-law distributions is easily seen. Norbert Wiener (I am a mathematician. *Science*, 1964) is said to have argued that 95% of the original works is made by less than 5% of all scientists. We examine some of the laws and use some concrete evidence from a study of the performance of a premier institute *X* (name withheld) with data from SCOPUS (www.scopus.com), an Elsevier product which is all set to become the single largest scientometric database with more than 27 million abstracts and citations covering 14,500 journals from 4000 publishers, and dating back to 1966. Free access to the beta version of this database made this study possible.

Methodology

The SCOPUS database was interrogated for all records of papers published by scientists from Institute *X* during 1974–2004. Here it is important to make the following distinctions. All databases will naturally indicate those individuals from the institute who have papers which are published in journals covered by the database. This set will vary according to how exclusive or inclusive the database is. SCOPUS is probably the largest scientometric database today and it is particularly generous to Indian journals, covering about 150 or so. We call all individuals (scientists) who appear in this list (i.e. who have at least one paper in the database for this period) as authors. Not all of the papers will have received citations (right now the database offers only citations obtained in the last ten years) and therefore many authors will remain uncited. There will also be a large number of the individuals (scientists) who will not have published during the period, or even if they have published, will not be fortunate to have their papers registered in the database because the journals in which they have published do not belong to the set of 14,500 journals covered by SCOPUS. Thus authors are a subset of scientists. Scientists without papers are called the zero item cases and for obvious reasons, their identities cannot be gleaned from the database. Sociologically, it is equally important to know who among the several hundreds that participate in the intellectual process associated with scientific discovery at Institute *X* never get to publish! This, of course, the scientometric data can never capture. For Institute *X*, in addition to all those who have achieved author status during the period covered by the study, an attempt is made to include all senior scientists (non-entry level, as it is assumed that entry-level scientists will take some time to get themselves established) on the latest roster as the zero item cases. For this study to be more complete, it is important that at a future date, the list of zero-paper scientists be obtained from the total population of all scientists who worked at Institute *X* during the period 1974–2004. For now, one must be satisfied with the present restriction, but it is felt that this will still give an indicative idea of the knowledge gathering and dissemination process at Institute *X*.

Lotka's Law

Figure 1 shows the histogram from a special arrangement of the scientometric data extracted from SCOPUS recently (early October 2004) and displayed in Table 1. It is important to note that the database is dynamic and changes every week! All papers from Institute *X* during 1974–2004 and all citations received since 1995 are registered. During this period, the number of unique authors, unique papers and whole count papers were: unique authors: 184; unique scientists: 366 and whole count papers: 882.

We use the whole count method for computation of the performance instead of fractionating a paper into partial counts. Thus, a paper with three authors will be counted as three whole counts or contributions. Accordingly, each author, whose name has appeared in a paper is given credit for the paper regardless of the number of co-authors. The data in Table 1 show (complete data available from the author upon request) that at the lower end, 182 scientists wrote no papers, 40 unique authors wrote only one paper each in the 30 years from 1974 to 2004, and another 52 authors contributed to two papers each. This distribution is captured in Figure 1. The distribution is bimodal, with a huge peak at 0 papers and another peak at 2 papers. This suggests that in Institute *X*, there is a large component of the task force doing only applied work (time-bound, mission-oriented, etc.) and does not publish, and that among those who publish, the peak is at 2 papers. That is, nearly 50% of the actual population of active scien-

Table 1. Actual number of scientists/authors

No. of papers	No. of scientists/ authors	Cum. whole count papers	% total whole count papers	Cum. scientist/ author count	% scientists/ authors
56	1	882	100.00	366	100.00
32	1	826	93.65	365	99.73
24	2	794	90.02	364	99.45
22	1	746	84.58	362	99.38
21	2	724	82.09	361	98.63
20	1	682	77.32	359	98.09
19	2	662	75.06	358	97.81
18	2	624	70.75	356	97.27
16	2	588	66.67	354	96.72
15	1	556	63.04	352	96.17
14	1	541	61.34	351	95.90
11	1	527	59.75	350	95.63
10	4	516	58.50	349	95.36
9	5	476	53.97	345	94.26
8	3	431	48.87	340	92.90
7	4	407	46.15	337	92.08
6	8	379	42.97	333	90.98
5	12	331	37.53	325	88.80
4	10	271	30.73	313	85.52
3	29	231	26.19	303	82.79
2	52	144	16.33	274	74.86
1	40	40	4.54	222	60.66
0	182	0	0.00	182	49.73
0	0	0	0.00	0	0.00

tists (a distinction which was made clear earlier) at Institute *X* produced no papers during this period. Among the 50% that has published, only a small fraction actually published 1 paper, and a larger number published 2 papers. When such distinctions are made, it becomes difficult to assess this evidence in the light of what Lotka[1] observed in 1926, where the population studied was all authors with names beginning with A or B in *Chemical Abstracts* covering the years 1907–16. About 60% of the authors produced only one paper during the period in Lotka's study. From this, he formulated his famous law of scientific productivity, whereby the number of authors making n contributions is about $1/n^2$ of those making 1.

However, the law is not accurate at the extreme tail of high-end scientific productivity. It is perfectly possible that we may find an author with a 100 papers and another with 50 or more papers. This is where Zipf's law comes to the rescue.

Zipf's law

At the higher end of the distribution, we see that the most productive authors produce much more than the average person. Indeed, we see from Table 1 that there are many individuals who have 10 (4 authors), 11, 14, 15, 16 (2 authors), 18 (2 authors), 19 (2 authors), 20, 21 (2 authors), 22, 24 (2 authors), 32, and 56 papers respectively. Zipf[2] was the first to record this and Zipf's law is one of rank frequency, which postulates that rank r occurs with a frequency which is inversely related to r. Note that a large number of variables are hidden in the system, but the rank-to-frequency relationship is captured in a simple way. Thus, if an author of the first rank has a 100 papers, an author of the second rank may have 50 (=100/2) or 25 (=100/2^2) papers, depending on the power of the inverse relationship. In this simple relationship that Zipf postulated, some kind of 'principle of least effort' was operating.

Lorenz curve and Pareto's law

The combined effect of the Lotka law at one end of the distribution and Zipf's law at the other end of the distribution is to confirm the general intuition described so well by Narin and Breitzman[3] that 'eminence is highly concentrated in a small fraction in the population', and that 'scientific creativity and productivity, are very highly concentrated in a population, and in the minds and abilities of a relatively small number of highly talented individuals'. Before we demonstrate that this is true for the present data, let us briefly review another variation that expresses this kind of inequality or disparity.

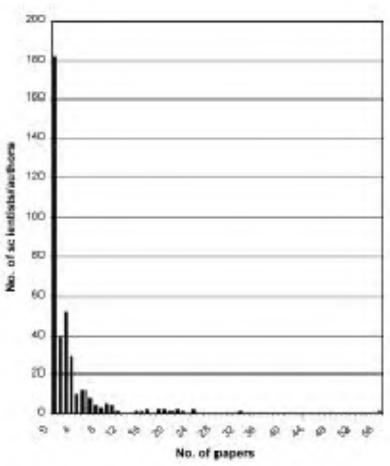

Figure 1. Histogram showing distribution of performance in terms of number of papers published.

The Italian engineer-turned-economist and political sociologist, Vilfredo Pareto realized that wealth is not evenly distributed[4]. Some of the people have most of the money. In fact, a fairly consistent minority, about 20% of people, controlled about 80% of a society's wealth. A closer examination would indicate that of the top 20% which owns 80% of the wealth, the 80–20 formula still applies reasonably consistently, so that the following pyramid can be set up as shown in Table 2. Thus, less than 1% or so of the population may account for 50% or so of the wealth. In Haiti, which has been in the news recently for all the wrong reasons, this is precisely the situation. In advanced countries like Australia, Japan and the United States, the top 1% accounts for 40% or more of the wealth.

That the same distribution is true for many other areas has been frequently noticed and is now termed the Pareto principle. Recently, press reports indicated that the three richest families in the world have as much wealth as the total population of the poorest 46 countries of the world!

We can display this inequality or disparity dramatically using a Lorenz curve, a device used by economists to represent inequality of income and wealth distribution in a population. In Table 1, columns 3 and 5 add up the whole count contributions and the authors cumulatively. Columns

Table 2. Pareto principle and the pyramid of wealth distribution

80–20 rule: 20% has 80%	0.800 has 0.200
80–20 rule on this 20% – $0.2 \times 0.2 = 0.04$ has $0.8 \times 0.8 = 0.64$	0.160 has 0.160
80–20 rule on this 4% – $0.2 \times 0.04 = 0.008$ has $0.8 \times 0.64 = 0.512$	0.008 has 0.512

so that

Pyramid of numbers	Pyramid of wealth
0.008	0.512
0.480	0.480
0.512	0.008

Table 3. Cumulative list of authors and citations

No. of citations	No. of authors	Cum. whole count citations	% total whole count citations	Cum. author count	% authors
199	1	1871	100.00	366	100.00
197	2	1672	89.36	365	99.73
126	1	1278	68.31	363	99.18
83	1	1152	61.57	362	98.91
66	1	1069	57.14	361	98.63
53	1	1003	53.61	360	98.36
47	1	950	50.77	359	98.09
43	1	903	48.26	358	97.81
40	1	860	45.96	357	97.54
34	2	820	43.83	356	97.27
32	1	752	40.19	354	96.72
31	1	720	38.48	353	96.45
30	1	689	36.83	352	96.17
22	2	659	35.22	351	95.90
20	1	615	32.87	349	95.36
19	2	595	31.80	348	95.08
18	3	557	29.77	346	94.54
17	2	503	26.88	343	93.72
16	2	469	25.07	341	93.17
15	2	437	23.36	339	92.62
14	1	407	21.75	337	92.08
13	1	393	21.00	336	91.80
12	2	380	20.31	335	91.53
11	3	356	19.03	333	90.98
10	0	323	17.26	330	90.16
9	6	323	17.26	330	90.16
8	5	269	14.38	324	88.52
7	7	229	12.24	319	87.16
6	7	180	9.62	312	85.25
5	8	138	7.38	305	83.33
4	5	98	5.24	297	81.15
3	12	78	4.17	292	79.78
2	11	42	2.24	280	76.50
1	20	20	1.07	269	73.50
0	249	0	0.00	249	68.03
0	0	0	0.00	0	0

4 and 6 show this in percentage terms. This is then displayed as a Lorenz curve in Figure 2. In Institute *X*, we find that 8% of the most productive scientists is responsible for 50% of the contributions. In the manner of Pareto's law, we see that 78% of the output is contributed by 22% of the authors (active scientists who publish one or more papers). One can interpret this to mean that the inequality in the distribution of scientific productivity at Institute *X* is only slightly less acute than indicated by the canonical 80:20 Pareto rule. This curve is incomplete in the sense that it does not incorporate the measure of the number of the scientists who were on the rolls of the Institute prior to 2003 but no longer in 2003 and who should have contributed during the 1974–2003 period but did not (i.e. the number of scientists who had 0 papers), for whatever reason. Only if this elusive group is identified will the Lorenz curve be complete. Perhaps, with this included, Pareto's 80:20 law also will be more closely approached.

To a science assessor, such distributions are valuable. About 10–20% of the scientists in any organization are responsible for about 50% of the papers published and the remaining 80–90% account for the remaining half of the output. Another 10–20% will be trying to move up the ladder (value chain in modern management terminology), and they should be encouraged. There will always be about 60% who will remain at the bottom.

The skewness of unfairness and scientific excellence

So far, we have dealt with the question of scientific productivity alone as evidenced by the number of publications in SCOPUS journals. Since these journals are chosen by some exclusive criteria, this is in itself a measure of quality for scientific output coming from a Third World country. We have seen from Lotka's and Zipf's laws and the Lorenz curve, how unfairly human ability is distributed across a given population, if the measure of ability is restricted to a raw count of the number of papers alone.

A popular measure of the quality of a paper is the number of citations that it has received in the open literature subsequent to publication. SCOPUS also provides a complete citation database for the last ten years of all papers published during the period 1974–2004. Again, we use the whole count method for computation of the performance instead of fractionating a citation into partial counts. Thus a citation for a paper with three authors will be counted as 3 whole counts of citation. Accordingly, each author, whose name has appeared in a paper is given credit for the citation regardless of the number of co-authors. Table 3 shows how the citations are distributed among the 366 authors. It is now possible to identify a group of hapless scientists (249 in number) who have not managed to collect a single cita-

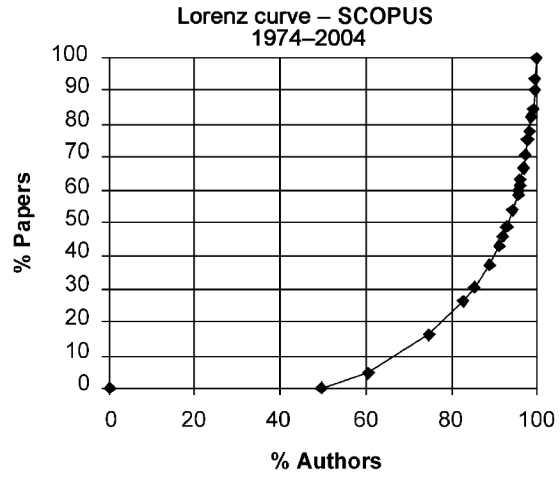

Figure 2. Lorenz curve shows that 8% of the high-end authors accounts for 50% of the output. In Pareto's terms, 78% of the output comes from the more productive 22% of the authors. Those with 0 papers have been identified, and introduced into the figure, and this helps explain why Pareto's 80: 20 law has been more closely approached.

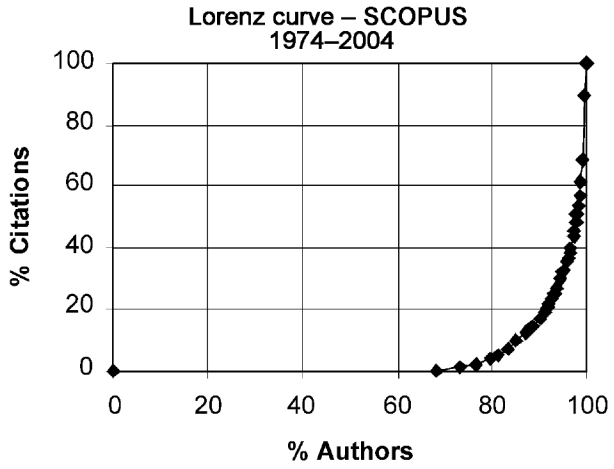

Figure 3. Lorenz curve for citations shows that less than 2% of the high-end authors accounts for 50% of the whole-count citations. The inequality is greater than the Pareto rule; in fact 12% of the authors accounts for 88% of the citations. This is because those with 0 citations have been identified; in fact 68% of the scientists have not received a citation during this period.

tion for any of their contributions over this 30-year period. This is almost 68% of all the scientists who appear in our list. At the lower end, it is seen, and frustratingly so, that Lotka's law could not be applied. One of the problems is the presence of this large group with 0 citations.

Similarly, at the higher end of performance, it was not easy to apply Zipf's law. However, it is clear from Table 3 that now approximately 2% of the high-end authors account for 50% of the whole-count citations. The inequality is greater than predicted by the canonical Pareto rule; in fact 12% of the authors account for 88% of the citations. This is because those with 0 citations have been identified, i.e. 68% of the scientists who have not received a citation during this period. This is graphically illustrated by the Lorenz curve in Figure 3. What we now see is that when scientific excellence is brought into the picture, the skewness of the unfairness is even more acutely emphasized.

Some sociological deconstruction

The main lesson from this exercise is that science assessment is too important a subject to be left to scientometricians who try to straight-jacket the data to simple formulae like Lotka's law and Zipf's law. We are dealing with complex situations governed by multiplicative random processes and hence Lorenz curves with such skewed tails are seen. We see from Institute *X*'s profile that there are two main groups. One serves the mission-oriented, time-bound projects, maybe about 90% of the population, with 0 or few papers. The smaller group does more academic and open-ended work. The existence of these two separate groups may explain the bimodal distribution seen clearly in the histogram in Figure 1.

We also see that even among the group that is presumably devoted to academic research, there is a high concentration of excellence in a small sub-group. Thus 2% of the scientists account for 50% of the highest quality work. However, such are the vagaries of the reward system that often the 2% of the scientists who account for 50% of the awards gathered are not from this deserving population. It is tempting to visualize a Pareto's law of luck, whereby 2% of the population has 50% of the luck! This underlines the need to have a more rational assessment procedure that will clearly demarcate the high quality workers and reward them accordingly.

1. Lotka, A. J., *J. Wash. Acad. Sci.*, 1926, **16**, 317–323.
2. Zipf, G., *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Cambridge, Massachusetts, 1949.
3. Narin, F. and Breitzman, A., *Res. Policy*, 1995, **24**, 507–519.
4. Pareto, V., *Cours d'Economie Politique*, Droz, Geneva, 1896.

Gangan Prathap is in the CSIR Centre for Mathematical Modelling and Computer Simulation, Bangalore 560 037, India
e-mail: gp@cmmacs.ernet.in