PERGAMON Mathematical and Computer Modelling 38 (2003) 323–329

MATHEMATICAL AND COMPUTER MODELLING

www.elsevier.com/locate/mcm

The Byline: Thoughts on the Distribution of Author Ranks in Multiauthored Papers

L. EGGHE
LUC, Universitaire Campus
B-3590 Diepenbeek, Belgium
and
UIA, IBW, Universiteitsplein 1
B-2610 Wilrijk, Belgium
leo.egghe@luc.ac.be

LIMING LIANG
Institute for Science, Technology and Society
Henan Normal University, Xinxiang, 453002, P.R. China
pllm@public.xxptt.ha.cn

R. ROUSSEAU
LUC, Universitaire Campus
B-3590 Diepenbeek, Belgium
and
UIA, IBW, Universiteitsplein 1
B-2610 Wilrijk, Belgium
ronald.rousseau@khbo.be

(Received and accepted February 2003)

Abstract—We analyze the multiauthorship matrix M, defined as the matrix where a cell $M(j,k)$ denotes the number of times authors with j publications are ranked as the k^{th} author of an article. We prove that if the distribution of the number of authors per paper follows a power law, then the author rank distribution is approximately equal to this power law (more precisely, equal in Landau's big O sense). We further determine the author rank distribution in the case where authors can be characterized through a seed number; this is the probability of preceding a fixed author in the byline of an article. Such a seed is determined for alphabetical ranking of authors using the standard western alphabet.

Keywords—Multiauthorship matrix, Byline, Author productivity distribution, Authors per paper, Rank distribution, Predicted rank, Seed.

INTRODUCTION

Laudel [1,2] defines research collaboration as a system of research activities by several actors related in a functional way, to attain a research goal corresponding with these actors' research

R. R. thanks the professors and students of the Institute for Science, Technology and Society, for their hospitality during his visit at the Henan Normal University, where part of the research for this article has been done. This research has been done in the framework of the Project 70073007 of National Natural Science Foundation of China.

0895-7177/03/$ - see front matter Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX
doi: 10.1016/S0895-7177(03)00223-1

goals or interests. Collaboration does not necessarily lead to a publication, or to coauthorship. In this article, however, we study from a structural, mathematical way how the final coauthorship relation of a whole group of scientists can be described and modeled.

In recent research of Chinese universities' scientific performance and collaboration structure, one of us (L.L.) encountered a matrix M (the multiauthorship matrix) of the following form: the element in cell (j,k), denoted as $M(j,k)$, represents the total number of times that authors with j publications are the k^{th} author (that is, occupies the k^{th} place in the byline of the publication). In this contribution, we investigate which structural elements one can derive from such a data matrix.

THE MULTIAUTHORSHIP MATRIX AND EMPIRICAL DISTRIBUTIONS

In order to clarify what we mean, we begin by presenting a small example. We consider a hypothetical database consisting of five articles. These articles are written by one, two, or three authors. The names of these authors are $X1$, $X2$, $X3$, and $X4$. Their names appear in the bylines as shown in Table 1.

Table 1. Illustrative database containing five articles.

Article 1	$X1$
Article 2	$X2$ - $X3$
Article 3	$X1$ - $X3$ - $X4$
Article 4	$X3$ - $X1$
Article 5	$X1$

The M-matrix for this database is shown in Table 2. The meaning of the symbols $R(k)$, $N(j)$, and $F(j)$ will be explained shortly.

Table 2. M-matrix corresponding to Table 1.

	Author Rank	1	2	3	$N(j)$	$F(j)$
Number of Articles	1	1	0	1	2	2
	2	0	0	0	0	0
	3	1	2	0	3	1
	4	3	1	0	4	1
	$R(k)$	5	3	1	9	4

Putting $R(k) = \sum_j M(j,k)$ yields the number of author-article pairs occurring at the k^{th} rank. As every article has exactly one first author (we assume that the database does not contain anonymous articles), $R(1) = T$ is equal to the total number of articles in the database. We also note that, on logical grounds, $R(k)$ must be a nonincreasing sequence. In our example, we see that $T = R(1) = 5$, $R(2) = 3$, and $R(3) = 1$. We denote by $M = \sum_{j,k} M(j,k)$ the total number of author-article pairs in the database (here $M = 9$). The sequence

$$r(k) = \frac{R(k)}{M}, \qquad k = 1, 2, \ldots \tag{1}$$

yields the discrete empirical rank distribution for authors in the database. Note that $r(1)$ is equal to T/M, the ratio of the total number of articles over the total number of entries in the matrix M. Its reverse: M/T is the average number of authors per article. Now put $N(j) = \sum_k M(j,k)$. Then $N(j)$ denotes the number of author-article pairs of authors that have authored (the case of

a single author, $j = 1$) or coauthored j articles (necessarily different ones!). The average rank of an author who has published j articles is then

$$\overline{R(j)} = \frac{\sum_k k\, M(j,k)}{N(j)}. \tag{2}$$

Dividing $N(j)$ by j gives the number of authors having j articles in the database (cf. Table 2),

$$F(j) = \frac{N(j)}{j}, \qquad j = 1, 2, \ldots. \tag{3}$$

Consequently, $F = \sum_j F(j)$ is the total number of (different) authors in the database ($F = 4$ in the example). The sequence

$$f(j) = \frac{F(j)}{F}, \qquad j = 1, 2, \ldots \tag{4}$$

is the discrete empirical distribution of articles per author. If we now denote by $k_{\max}$ the largest number of authors in one article ($k_{\max} = 3$ in the example), then $R(k)$ takes values for $k = 1 \to k_{\max}$. Putting $A(k)$ equal to the number of articles with k authors, we obtain the following relation between $A(k)$ and $R(k)$:

$$A(k_{\max}) = R(k_{\max}) \quad \text{and} \quad A(k) = R(k) - R(k+1), \qquad k = 1, \ldots, k_{\max} - 1. \tag{5}$$

Note that $T = R(1) = \sum_{k=1}^{k_{\max}} A(k)$. For the example, we obtain $A(3) = 1$, $A(2) = 3 - 1 = 2$, and $A(1) = 5 - 3 = 2$. The discrete, empirical distribution of authors per article is given as

$$a(k) = \frac{A(k)}{T}, \qquad k = 1, \ldots, k_{\max}. \tag{6}$$

The relation between the discrete distribution of authors per article and the rank distribution is given by

$$a(k_{\max}) = \frac{r(k_{\max})}{r(1)} \quad \text{and} \quad a(k) = \frac{M}{T}\left(r(k) - r(k+1)\right), \qquad k = 1, \ldots, k_{\max} - 1. \tag{7}$$

Similarly, we see that

$$\begin{aligned} R(k) &= A(k) + R(k+1) \\ &= A(k) + A(k+1) + R(k+2) \\ &= \cdots = \sum_{i=k}^{k_{\max}} A(i), \end{aligned} \tag{8}$$

a relation which also holds for $k = k_{\max}$. For the corresponding discrete distributions, we have for $k = 1, \ldots, k_{\max}$,

$$r(k) = \frac{T}{M} \sum_{i=k}^{k_{\max}} a(i).$$

The distributions $f(j)$, the author productivity distribution, i.e., articles per author, and $a(k)$, the discrete byline density distribution, i.e., authors per article, are said to be each other's dual [3]. All this shows that the multiauthorship matrix M has a rich structure.

A GENERAL MODEL FOR THE AUTHOR RANK DISTRIBUTION

We denote by $r(k \mid m)$ the conditional rank distribution given that a paper has m authors, and assume that this distribution is the uniform one: $r(k \mid m) = 1/m$. By this we mean that the probability that a particular author occupies the k^{th} rank, given that the article has m authors, is $1/m$. So, any rank is equally probable. Then we have the following proposition.

PROPOSITION A. *The author ranking distribution, $r(k)$, is given by*

$$r(k) = \sum_{m=k}^{\infty} \frac{a(m)}{m}. \tag{9}$$

PROOF. The result follows immediately. Indeed,

$$r(k) = \sum_{m=k}^{+\infty} r(k \mid m)a(m) = \sum_{m=k}^{+\infty} \frac{a(m)}{m}.$$

We have applied here a basic result about conditional probabilities (see, e.g., [4, p. 116]). It is sometimes referred to as the theorem of total probability.

We recall the following notation, originally due to the German mathematician Landau.

DEFINITION. *(See [5].) Consider two sequences $a(n)_n$ and $b(n)_n$. One writes that $a(n) = O(b(n))$ if there exist numbers n_0 and C such that, for $n \geq n_0$,*

$$|a(n)| \leq C|b(n)|.$$

Intuitively, this means that the sequence $a(n)_n$ does not grow faster than the sequence $b(n)_n$. This notation leads to an elegant formulation of the next theorem.

THEOREM B. *If the distribution of numbers of authors per article, $a(m)$, is given by a Lotka distribution, then the author ranking distribution, $r(m)$, is related to $a(m)$ by*

$$r(m) = O(a(m)). \tag{10}$$

PROOF. If $a(m) = C/m^{\alpha}$, $\alpha > 0$, then, by Proposition A, $r(m) = \sum_{n=m}^{+\infty}(C/n^{\alpha+1})$. Consequently, by the integral test and the fact that the series $\sum_m(C/m^{\alpha+1})$ is convergent [5], we have

$$\sum_{n=m}^{+\infty} \frac{C}{n^{\alpha+1}} - \int_m^{+\infty} \frac{C}{x^{\alpha+1}}\,dx \leq \frac{C}{m^{\alpha+1}}.$$

Hence,

$$\sum_{n=m}^{+\infty} \frac{C}{n^{\alpha+1}} - \frac{C}{\alpha m^{\alpha}} = r(m) - \frac{C}{\alpha m^{\alpha}} \leq \frac{C}{m^{\alpha+1}}$$

or

$$r(m) \leq \frac{C}{m^{\alpha}}\left(\frac{1}{m} + \frac{1}{\alpha}\right) \leq \frac{C}{m^{\alpha}}\left(1 + \frac{1}{\alpha}\right) = a(m)\left(1 + \frac{1}{\alpha}\right).$$

This proves the theorem.

We are aware of the fact that real author-rank distributions rarely follow a Lotka distribution [6–8], but, as done in other publications, we use this model as a first approximation [9,10].

MODELING THE AUTHOR RANK DISTRIBUTION USING SEEDS

Assume that each author, denoted here as A, has a characteristic number $s_A \in [0, 1]$. This characteristic number, s_A, is equal to the probability that another author comes before A in the byline of an article. The number s_A will be called a 'seed' and it characterizes the 'average' position of author A in the byline of an article.

We will next solve the problem of determining $r(k, s)$: the probability for an author with seed number s to be the k^{th} author (in general); or more specifically $r(k, s \mid m)$, the probability of an author with seed s to be the k^{th} author in a publication with m authors. In this connection, we have the following result.

PROPOSITION C.

$$r(k, s \mid m) = \binom{m-1}{k-1} s^{k-1}(1-s)^{m-k}, \tag{11}$$

$$r(k, s) = \sum_{m=k}^{\infty} \binom{m-1}{k-1} s^{k-1}(1-s)^{m-k} a(m), \tag{12}$$

where $a(m)$ denotes the probability that a paper has m authors.

PROOF. Consider an author A, with seed s_A. Since s_A is author A's seed, we know that

$$P(\text{an author is before } A \text{ in an author list}) = s,$$
$$P(\text{an author is after } A \text{ in an author list}) = 1 - s.$$

Author A has rank k in an article with m authors, m being at least equal to k, if and only if $k-1$ authors precede A, and $m-k$ follow A. We can describe this as follows. As the article has m authors, this means that $m-1$ coauthors are chosen at random. They end up before A with probability s (we refer to this as 'success' in a Bernoulli trial). So author A ends up at rank k if there are $k-1$ successes (and consequently, $m-k$ 'failures'). This shows that the situation can be described by a binomial distribution

$$r(k, s \mid m) = \binom{m-1}{k-1} s^{k-1}(1-s)^{m-k}. \tag{11}$$

The second formula, where the total number of authors of an article is not given, follows by the law of total probability:

$$r(k, s) = \sum_{m=k}^{\infty} r(k, s \mid m) a(m) \tag{13}$$

$$= \sum_{m=k}^{\infty} \binom{m-1}{k-1} s^{k-1}(1-s)^{m-k} a(m). \tag{12}$$

This proves Proposition C.

FINDING A SEED BASED ON ALPHABETICAL RANKING OF AUTHORS

In this section, we introduce a method of finding a seed for an author. First, however, we define an injection (a one-one relation) between an author's name and a rational number in the set [0,1].

We will work with the standard western alphabet, consisting of 26 letters, but the method applies to any other alphabet consisting of symbols with a fixed rank. We add a 0 symbol to the alphabet, so that we have an alphabet of 27 symbols. Let **S** denote the set of all concatenations of a finite or infinite number of symbols. Then $S_1 \ldots S_n$ (a concatenation of a finite number of nonzero symbols) represents an arbitrary name. The injection

$$f : \mathbf{S} \to [0, 1]$$

is defined as

$$f(S_1 S_2 \ldots S_n) = \sum_{i=1}^{n} \frac{|S_i|}{27^i}, \tag{14}$$

where $|S_i|$ denotes the rank of the symbol S_i. An equivalent way of defining the function f is

$$f(S_1 S_2 \ldots S_n) = 0.|S_1| \ldots |S_n|,$$

where $0.|S_1|\ldots|S_n|$ denotes a number in the 27-ary number system. For clarity's sake, each number $|S_j|$ must be expressed by two digits, otherwise a 1 followed by a 2 could be confused with 12. Hence, 1 must be written as 01, 2 as 02, and so on.

EXAMPLES.

(1) $f(A) = 0.|A| = 1/27 = 27^{-1}$.

(2)

$$f(AZZZZZZZZZZZ\ldots)$$
$$= \frac{1}{27} + \sum_{i=2}^{\infty} \frac{26}{27^i} = \frac{1}{27} + \frac{26}{27^2(1-1/27)} = \frac{1}{27} + \frac{1}{27} = \frac{2}{27} = 0.|B| = f(B).$$

This second example shows that, in the same way as we identify $0.19999\ldots$ with $0.2 \in \mathbb{R}$, $AZZZZZZ\ldots$ is identified with B.

(3)
$$f(ZZZZ\ldots) = \sum_{i=1}^{\infty} \frac{26}{27^i} = \frac{26}{27(1-1/27)} = 1.$$

(4) Note that $f(000\ldots0*000\ldots) = \text{rank}(*)/27^i$, where $*$ denotes any symbol from the alphabet, placed after $(i-1)$ zeros. Of course, this symbol does not represent an existing name. Clearly, the limit of this expression, for $i \to \infty$, is zero.

(5) The image of any existing name belongs to $[0,1] \cap \mathbb{Q}$.

The function f, restricted to 'possibly existing names' (the subset of finite symbols of the form $S_1\ldots S_n$), is a one-one relation. Indeed, let

$$f(S_1S_2\ldots S_n) = f(T_1T_2\ldots T_m).$$

Hence,

$$\sum_{i=1}^{n} \frac{|S_i|}{27^i} = \sum_{j=1}^{m} \frac{|T_j|}{27^j}.$$

Assume now that $S_1\ldots S_n \neq T_1\ldots T_m$ (that is, assume that f is not an injection). Let $k \in \{1,\ldots,\min(m,n)\}$ be the first rank for which $S_k \neq T_k$. There is no loss in generality in assuming that $|S_k| \geq |T_k| + 1$. Hence,

$$\frac{|S_k|}{27^k} - \frac{|T_k|}{27^k} \geq \frac{1}{27^k}. \tag{15}$$

But, we always have that

$$\sum_{j=k+1}^{m} \frac{|T_j|}{27^j} - \sum_{j=k+1}^{n} \frac{|S_j|}{27^j} < \sum_{j=k+1}^{\infty} \frac{26}{27^j} = \frac{26}{27^{k+1}} \frac{1}{1-1/27} = \frac{1}{27^k}, \tag{16}$$

which is in contradiction with (15) and the fact that $f(S_1S_2\ldots S_n) = f(T_1T_2\ldots T_m)$. Hence, the function f is an injection on the subset of 'existing names'.

How can the actual occurrence of letters be taken into account? One suggestion is to use a telephone directory. This suggestion, however, only works for scientists 'speaking' the same language, because the use of letters differs in different languages. Otherwise, one needs an 'international' directory (perhaps that of a city such as New York or Los Angeles). Then a name listed on page 345 of the 1557 (this is just an example) would get a seed equal to 345/1557 = 0.22158. Further refinements (using lines within a page) are possible. Averaging would be necessary for popular names.

Comments

It can be tested whether the distributions proposed in Proposition C correspond with reality by using a group of scientists with the same surname (hence, the same seed) in a field where alphabetic ranking of authors is customary. Such is generally the case for pure mathematicians, logicians, statisticians, and theoretical physicists [11]. Proposition C predicts the rank distribution of such an author.

This model is valid in all cases where a seed can be given, not just in the case of alphabetical name ordering. Indeed, there exist many ways and conventions for ranking coauthors [11–13]. A seed can, for instance, be derived from the importance of the author. Indeed, assume that author ranking always occurs according to 'importance'. 'Importance' could (just as an example) be obtained from the number of publications, the number of citations, or even the age of scientists [14].

Another suggestion is to calculate a seed from 'older' publications (calculating an average rank) and to use this to 'predict' the author rank distribution in 'newer' ones.

Note that a seed, being a probability, is always a number in the interval [0,1], so that observations must always be transformed to the unit interval in order to obtain a seed.

CONCLUSIONS

We have analyzed the multiauthorship matrix, explaining different relations and statistical distributions that can be derived from such a matrix representation. Next we have modeled the multiauthorship relation based on the notion of a seed. More specifically, we found that if the distribution of the number of authors per paper follows a Lotka distribution, the distribution of author ranks follows a Lotka distribution too, at least in the O-sense. Finally, introducing the 27-ary number system, we showed how such a seed can be obtained for the standard western alphabet and alphabetic ranking of coauthors.

REFERENCES

1. G. Laudel, Collaboration, creativity and rewards: Why and how scientists collaborate, *International Journal of Technology Management* **22** (8), 762–781, (2001).
2. G. Laudel, What do we measure by co-authorships?, In *Proceedings of the 8^{th} International Conference on Scientometrics and Informetrics*, (Edited by M. Davis and C. Wilson), pp. 369–384, BIRG (UNSW), Sydney, (2001).
3. L. Egghe, The duality of informetric systems with applications to the empirical laws, *Journal of Information Science* **16**, 17–27, (1990).
4. W. Feller, *An Introduction to Probability Theory and Its Applications, Volume I*, Third Edition, Wiley, New York, (1967).
5. T. Apostol, *Mathematical Analysis*, Addison-Wesley, Reading, MA, (1974).
6. I. Ajiferuke, A probabilistic model for the distribution of authorships, *Journal of the American Society for Information Science* **42**, 279–289, (1991).
7. B.M. Gupta and R. Rousseau, Further investigations into the first-citation process: The case of population genetics, *LIBRES: Library and Information Research Electronic Journal* **9** (2), (1999); `http://aztec.lib.utk.edu/libres/libre9n2/fc.ftm`.
8. R. Rousseau, The number of authors per article in library and information science can often be described by a simple probability distribution, *Journal of Documentation* **50**, 134–141, (1994).
9. L. Egghe, Consequences of Lotka's law in the case of fractional counting of authorship and of first author counts, *Mathl. Comput. Modelling* **18** (9), 63–77, (1993).
10. L. Egghe and I.K.R. Rao, Duality revisited: Construction of fractional frequency distributions based on two dual Lotka laws, *Journal of the American Society for Information Science and Technology* **53**, 789–801, (2002).
11. M.A. Harsanyi, Multiple authors, multiple problems—Bibliometrics and the study of scholarly collaboration: A literature review, *Library and Information Science Research* **15**, 325–354, (1993).
12. W.P. Hoen, H.C. Walvoort and A.J.P. Overbeke, What are the factors determining authorship and the order of the authors' names?, *Journal of the American Medical Association* **280** (3), 217–218, (1998).
13. H. Feger, Co-authorship patterns in work reports, Presented at the *Second COLLNET Workshop on Scientometrics and Informetrics*, Hohen Neuendorf, September 1–4, 2000.
14. L. Liang, H. Kretschmer, Y. Guo and D. DeB. Beaver, Age structures of scientific collaboration in Chinese computer science, *Scientometrics* **52**, 471–486, (2001).