The Byline: Thoughts on the distribution of author ranks in multiauthored papers

Non Peer-reviewed author version

EGGHE, Leo; Liming, Liang & ROUSSEAU, Ronald (2003) The Byline: Thoughts on the distribution of author ranks in multiauthored papers. In: Mathematical and Computer Modelling, 38(3-4). p. 323-329.

DOI: 10.1016/S0895-7177(03)90090-2
Handle: http://hdl.handle.net/1942/766

The byline: thoughts on the distribution of author ranks in multi-authored papers

Leo Egghe [1] , Liming Liang [2], Ronald Rousseau [1]

[1] LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium
and UIA, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium
e-mail: e-mail: leo.egghe@luc.ac.be ; ronald.rousseau@khbo.be

[2] Henan Normal University
Institute for Science, Technology and Society
Xinxiang, 453002, P.R. China
e-mail: pllm@public.xxptt.ha.cn

Abstract

We analyze the multi-authorship matrix M, defined as the matrix where a cell M(j,k) denotes the number of times authors with j publications are ranked as 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 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: multi-authorship matrix, byline, author productivity distribution, authors per paper, rank distribution, predicted rank, seed, 27-ary number system

Introduction

Grit Laudel [1,2] recently 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 goals or interests. Collaboration does not necessarily lead to a publication, nor to co-authorship. In this article, however, we study, from a structural, mathematical way how the final co-authorship relation of a whole group of scientists can be described and modelled.

In recent research of Chinese universities' scientific performance and collaboration structure one of us (L.L.) encountered a matrix M (the multi-authorship 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 k^{th} author (this 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 multi-authorship matrix and empirical distributions

In order to clarify what we mean we begin by presenting a small example. We consider an hypothetical database consisting of five articles. These articles are written by 1, 2 or 3 co-authors. The names of these authors are X1, X2, X3 and X4. Their names appear in the bylines in the following order:

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 then:

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

The meaning of the symbols R(k), N(j) and F(j) will be explained shortly.

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 non-increasing 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

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

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. Put now 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 co-authored j articles (necessarily different ones!). The average rank of an author who has published j articles is then

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

Dividing N(j) by j gives the number of authors having j articles in the database:

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

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

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

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):

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

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:

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

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

$$a(k_{\max}) = \frac{r(k_{\max})}{r(1)}$$

and

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

Similarly, we see that

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

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 each other's dual [3].

A general model for the author rank distribution

We denote by $r(k|m)$ the conditional rank distribution given that a paper has m authors, and assume that this distribution is the uniform one: $r(k|m)$ = 1/m. Then we have the following proposition.

Proposition A

The author ranking distribution, r(k), is given by

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

Proof.

The result follows immediately from the theorem of total probability. Indeed:

$$r(k)=\sum_{m=k}^{+\infty} r\left(k\middle|m\right)a(m)=\sum_{m=k}^{+\infty}\frac{a(m)}{m}$$

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

Definition [4]

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:

$$r(m)=O(a(m)) \qquad (10)$$

Proof.

If $a(m) = \frac{C}{m^{\alpha}}$, $\alpha > 0$, then, by proposition A, $r(m) = \sum_{n=m}^{+\infty}\frac{C}{n^{\alpha+1}}$. Consequently, by the integral test and the fact that the series with $\frac{C}{m^{\alpha+1}}$ as terms is convergent, 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 [5-7], but, as in other publications, we use this model as a first approximation, cf. [8,9].

Modelling the author rank distribution using seeds

Assume that each author, A, has a characteristic number $s_A \in [0,1]$, where s_A is equal to the probability that an other author comes before A in the byline of an article. This characteristic number will be called a 'seed'.

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 specific r(k,s|m): the probability of an author with seed s to be k^{th} author in a publication with m authors. In this connection we have the following result.

Proposition C

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

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

where a(m) is 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(an author is before A in an author list) = s

P(an author is after A 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 co-authors 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.

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

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

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

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

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 we will introduce an injection between an author's name and a number in the set [0,1] $\cap \mathbb{Q}$.

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 non-zero symbols) represents an arbitrary name. The injection

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

is defined as:

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

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 $|S_i|$ 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) = $0 \cdot |A| = \frac{1}{27} = 27^{-1}$

2) f(AZZZZZZZZZZ .....) =

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

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

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

4°) Note that, $f(000...0*000....)=\frac{rank(*)}{27^i}$, where * denotes any symbol from the alphabet, placed after (i-1) zeros. Of course, this symbol does not represent a real name. Clearly, the limit of this expression, for $i\to\infty$, is zero.

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

The function f, restricted to 'real names', (the subset of finite symbols of the form $S_1...S_n$,) is an injection. Indeed, let

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

Hence,

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

Assume now that $S_1...S_n \neq T_1...T_m$, (this is: assume that f is not an injection). Let k $\in\{1,...,\min(m,n)\}$ 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,

$$\frac{|S_k|}{27^k} - \frac{|T_k|}{27^k} \geq \frac{1}{27^k} \qquad (15)$$

But, we always have that

$$\sum_{j=k+1}^{m} \frac{|T_j|}{27^j} - \sum_{j=k+1}^{n} \frac{|S_j|}{27^j} < \sum_{j=k+1}^{\infty} \frac{26}{27^j} = \frac{26}{27^{k+1}} \frac{1}{1-\frac{1}{27}} = \frac{1}{27^k} \qquad (16)$$

which is in contradiction with (15) and the fact that $f(S_1S_2...S_n) = f(T_1T_2...T_m)$. Hence the function f is an injection on the subset of 'real 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 [10]. 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 co-authors ([10-12]). A seed can, e.g., be derived from the importance of the author. Indeed, assume that author ranking always occurs according to 'importance'. 'Importance' could then be calculated from the number of publications, the number of citations, or even the age of scientists [13].

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 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 analysed the multi-authorship matrix, making clear different relations and distributions that can be derived from such a matrix representation. Next we have modelled the multi-authorship 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 show how such a seed can be obtained for the standard western alphabet and alphabetic ranking of co-authors.

Acknowledgements

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.

References

[1] G. Laudel, Collaboration, creativity and rewards: why and how scientists collaborate, *International Journal of Technology Management* 22(8), (2001) 762-781.

[2] G. Laudel, What do we measure by co-authorships? In: M. Davis & C. Wilson (eds.), *Proceedings of the 8th International Conference on Scientometrics & Informetrics* (BIRG (UNSW), Sydney (Australia) 2001) 369-384.

[3] L. Egghe, The duality of informetric systems with applications to the empirical laws, *Journal of Information Science* 16 (1990) 17-27.

[4] T. Apostol, *Mathematical analysis* (Reading (MA), Addison-Wesley, 1974).

[5] 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 (1994) 134-141.

[6] 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

[7] I. Ajiferuke, A probabilistic model for the distribution of authorships, *Journal of the American Society for Information Science*, 42 (1991) 279-289.

[8] L. Egghe, Consequences of Lotka's law in the case of fractional counting of authorship and of first author counts, *Mathematical and Computer Modelling* 18 (1993) 63-77.

[9]. L. Egghe and I.K. Ravichandra Rao, Duality revisited: construction of fractional frequency distributions based on two dual Lotka laws, Preprint (2001).

[10] M.A. Harsanyi, Multiple authors, multiple problems – Bibliometrics and the study of scholarly collaboration: a literature review, *Library and Information Science Research* 15 (1993) 325-354.

[11] W.P. Hoen, H. C. Walvoort, and A.J.P.M. Overbeke, What are the factors determining authorship and the order of the authors' names? *Journal of the American Medical Association*, 280(3) (1998) 217-218.

[12] H. Feger, Co-authorship patterns in work reports. Paper presented at the Second COLLNET Workshop on Scientometrics and Informetrics, September 1-4, 2000. Hohen Neuendorf.

[13] L. Liang, H. Kretschmer, Y. Guo and D. DeB. Beaver, Age structures of scientific collaboration in Chinese computer science, *Scientometrics* 52 (2001) 471-486.