Empirical and combinatorial study of country occurrences in multi-authored papers

Empirical and combinatorial study of country occurrences in multi-authored papers

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium

leo.egghe@uhasselt.be

ABSTRACT

Papers written by several authors can be classified according to the countries of the author affiliations. The empirical part of this paper consists of two datasets. One dataset consists of 1,035 papers retrieved via the search "pedagog∗" in the years 2004 and 2005 (up to October) in Academic Search Elite which is a case where $\varphi(m)$= the number of papers with $m = 1,2,3,...$ authors is decreasing, hence most of the papers have a low number of authors. Here we find that $\#_{j,m}$ = the number of times a country occurs j times in an m-authored paper,

---

[1] Permanent address

Key words and phrases: country occurrence, multi-authored paper.
Acknowledgement. The author is indebted to Drs. M. Goovaerts for the "readily" preparation of the list of papers derived from the IR process "pedagog∗" (2004 + 2005 until October) and "enzym∗" (2005 until October) in EBSCO's Academic Search Elite.

$j = 1,...,m-1$ is decreasing and that $\#_{m,m}$ is much higher than all the other $\#_{j,m}$ values. The other dataset consists of 3,271 papers retrieved via the search "enzym*" in the year 2005 (up to October) in the same database which is a case of a non-decreasing $\varphi(m)$: most papers have 3 or 4 authors and we even find many papers with a much higher number of authors. In this case we show again that $\#_{m,m}$ is much higher than the other $\#_{j,m}$ values but that $\#_{j,m}$ is not decreasing anymore in $j = 1,...,m-1$, although $\#_{1,m}$ is (apart from $\#_{m,m}$) the largest number amongst the $\#_{j,m}$.

The combinatorial part gives a proof of the fact that $\#_{j,m}$ decreases for $j = 1,...,m-1$, supposing that all cases are equally possible. This shows that the first dataset is more conform with this model than the second dataset. Explanations for these findings are given.

From the data we also find the (we think: new) distribution of number of papers with $n = 1,2,3,...$ countries (i.e. where there are n different countries involved amongst the $m(\geq n)$ authors of a paper): a fastly decreasing function e.g. as a power law with a very large Lotka exponent.

I. Introduction

One of the fascinating topics in informetrics is multiple authorship in documents (e.g. papers). In terms of sources and items (see e.g. Egghe and Rousseau (1990), Egghe (2005a)) we have here a case where items can have multiple sources and hence one has here a "counting" problem since it is an interesting debate how to count the author credits in such multi-authored papers – see e.g. Egghe (1993), Egghe, Rousseau and Van Hooydonk (2000), Egghe and Rao (2002), Gauffriau and Larsen (2005), Chapter VI in Egghe (2005a), Harsanyi (1993), Kranakis and Kranakis (1988), Lindsey (1980). To the best of our knowledge, the study of items with multiple sources is a typical informetrics topic which is not really encountered in other –metrics fields such as econometrics or biometrics. The determination of the counting method is also decisive for author performance rankings and these rankings are quite different from one counting method to another one (see Kranakis and Kranakis (1988), Egghe,

Rousseau and Van Hooydonk (2000), Gauffriau and Larson (2005) and Chapter VI in Egghe (2005a)).

Multiple authorship in papers also expresses an aspect of duality in informetrics: usually one is interested in authors (as sources) and their number of papers they have written (as items) but in the case of multiple authorship one can also consider papers (as sources) having a certain number of authors (as items). In this connection informetrics researchers have conducted research in order to see whether the classical law of Lotka (Lotka (1926)), describing the number of authors with a certain number of papers, is also valid for "number of papers with a certain number of authors". An overview of papers on this topic is given in Chapter VI in Egghe (2005a) but here we can, briefly, state that – essentially – there are two types of paper-authorship relations: one more or less conforming with the decreasing Lotka power law (i.e. the number of papers with $m = 1,2,3,...$ authors is a decreasing function) and one where the mode of the function is obtained for a certain value of number of authors $m > 1$ (i.e. the number of papers with $m = 1,2,3,...$ authors is fastly increasing e.g. to $m = 3$ or 4 or 5 … and then slowly decreasing for larger values of m). The two datasets analyzed in this paper show this different nature. One dataset covers the topic "pedagog∗" in EBSCO's Academic Search Elite (2004 + 2005 up to October) and shows the decreasing relationship for $\varphi(m)$= the number of papers with $m = 1,2,3,...$ authors. The other dataset covers the topic "enzym∗" (also in Academic Search Elite) in the period 2005 (up to October) and shows a fast increase of $\varphi(m)$ for $m = 1,2,3$ and then a slow decrease from $m = 4$ onwards.

Although it is not the main topic of this paper, in Section II, we will analyse the function $\varphi(m)$ in both cases in the next section. We will also derive, from our data (both datasets), the new function $\psi(n)$ being the number of papers with n countries, i.e. the number of papers which have been written by authors representing (in total) n countries (where we deleted papers with authors which indicated more than one author affiliation in different countries). Contrary to the $\varphi(m)$ function, now the function $\psi(n)$ is always decreasing and in fact very fastly decreasing. Details will also be given in Section II.

In Section III we present data on the numbers $\#_{j,m}, j = 1,2,...,m$ being the number of times a country appears j times in a paper with m authors. A first, intuitive idea, suggests that the

sequence $\#_{1,m}, \#_{2,m}, ..., \#_{m,m}$ should be increasing since the probability to collaborate between authors of the same country should be highest. This last idea is certainly true and is reflected in the fact that – in both datasets – for all $m \geq 2$, $\#_{m,m}$ is the highest value amongst the $\#_{1,m}, ..., \#_{m,m}$ and we even have

$$\#_{m,m} \geq \#_{j,m} \quad (1)$$

for all $j = 1, ..., m-1$, indicating that the majority of the multi-authored papers is written by authors coming from the same country. But, in the study of $\#_{j,m}$ for $j = 1, ..., m-1$, there is also a combinatorial aspect involved. It is indeed so that most collaboration is done within the same country but this also implies that there should be a high probability for $m-1$ authors (out of the m authors of a paper) to come from the same country (higher than for $m-2, m-3, ...$ authors from the same country). But this immediately implies that there is also a country that occurs only once as affiliation of an author. In other words, if there are m authors of a paper the occurrence of a country representing $m-1$ authors implies a score 1 for another country. Also if there are $m-2$ authors from the same country this implies that a score 2 or <u>two times</u> a score 1 appears (if the two other authors are from the same, respectively a different, country) and so on.

This intuitive argument shows that, purely for combinatorial (i.e. non-informetric) reasons the number 1 as an occurrence of a country should be found oftenly and hence that the number $\#_{1,m}$ should be relatively high in the list $\#_{1,m}, ..., \#_{m-1,m}$ (but smaller than $\#_{m,m}$). In the same way one could expect (intuitively) that the number 2 appears quite often (but less than the number 1) and more than the numbers $3, 4, ..., m-1$. In other words we have here an intuitive argument for the following regularity, for all $m \geq 2$

$$\#_{m,m} \geq \#_{1,m} > \#_{2,m} > \#_{3,m} > ... > \#_{m-1,m} \quad (2)$$

The data on "pedagog*" indeed show this regularity for all m for which we have enough papers (as said above, in this set the number of co-authors is low and papers with 6 or more authors are hardly existing, hence the determination of the $\#_{j,m}$ is not very meaningful in these cases).

It is for this reason that we also analyzed the large dataset "enzym∗", a topic where one has much more collaboration, as is well-known (and indicated above). Here we were able to analyse $\#_{j,m}\ (j=1,...,m)$ for many values of m: in fact up to $m=12$ (included) we had enough data to see the following regularity: (2) is not valid anymore but we have

$$\#_{m,m} ?\ \#_{1,m} > \#_{j,m} \tag{3}$$

for all $j=2,...,m-1$, but the order between the values $\#_{2,m},\#_{3,m},...,\#_{m-1,m}$ is not very clear.

We think the above sheds some light on this very new distribution of the $\#_{j,m}\ (j=1,...,m)$. In Section IV we will prove that the sequence $(\#_{j,m})_{j=1,...,m}$ is decreasing, supposing that all partitions of m as country occurrences are equally possible. In practise this is not true and the argument overestimates the 1-occurrences and of course underestimates $\#_{m,m}$ (see (2),(3)) but, nevertheless, it yields a first, basic, explanation of (2) as found in "pedagog∗" and of (3) as found in "enzym∗".

The paper closes with conclusions and some open problems.

II. Data collection and study of the functions $\varphi(m)$= number of papers with m authors and $\psi(n)$= number of papers with n countries

II.1 First dataset

The first dataset that was constructed was the result of a retrieval process in EBSCO's Academic Search Elite with command "pedagog*" in the period 2004 and 2005 (up to October). We checked, for each document, the number of authors and their country affiliations. After removal of the cases with authors with two or more affiliations in different countries and of the cases with an affiliation that was unclear (in the database indexation) we ended up with 1,937 documents. Removal of the cases where an author belongs to more than one country was the only solution: if such affiliations are counted in a total way this biases the number of authors and of countries per paper; if such affiliations are counted fractionally we get fractional values of $j \in [0,m]$ for $\#_{j,m}$ = the number of times a country appears $j \in \mathbb{N}^+$ times in a paper with m authors: this is a complication we want to avoid since the "simple" case of $\#_{j,m}$ for $j = 1,...,m$ is new and far from trivial (fractional author credits were studied in Egghe and Rao (2002) from which the complexity is very clear). This deletion of papers with multi-country author affiliations might cause a small bias in our sample (perhaps we delete more papers with a high number of authors – but this is not sure) but we are convinced this will not jeopardise our results fundamentally (also because we encountered only 6 of such papers). To be clear: it is evident that papers with authors with more than one affiliation in the same country were kept in our sample and the country occurrence for this author was counted as one. Finally, if there was only one author, the paper was not excluded from the sample since it adds (correctly) to the case $m = 1$ (the number $\varphi(1)$ of papers with one author). Of course, for $m = 1$, the numbers $\#_{j,m}$ are not interesting (only $\#_{1,1}$ exists).

Each paper in this sample was added to a table "$m = 1,2,3,...$" and denoted $|j_1 j_2 ... j_k|$ where, if there are m authors,

$$\sum_{i=1}^{k} j_i = m \qquad (4)$$

and where j_i denotes the number of times a certain country appears as affiliation of an author of this paper. Example: $|431|$ means a paper with 8 authors, four of them of the same country,

3 of them of another country (the same for these 3 authors) and, finally, 1 author with yet another country affiliation.

We got the following results. The empirical $\varphi(m)$= number of papers with m authors was as in Table 1.

Table 1. Number of papers with m authors (for "pedagog∗").

m	$\varphi(m)$ (empirical)	fraction: $\frac{\varphi(m)}{1{,}937}$
1	1,287	0.6644
2	367	0.1895
3	152	0.0785
4	61	0.0315
5	37	0.0191
6	13	0.0067
7	4	0.0021
8	2	0.0010
9	5	0.0026
10	1	0.0005
11	5	0.0026
12	1	0.0005
13	2	0.0010

According to the table of $\frac{1}{\zeta(\alpha)}$ (Egghe and Rousseau (1990), Table IV.6.6., p. 357 or Table A.1, p. 384 in Egghe (2005a)) we can expect a Lotka value of $\alpha = 2.18$ found as follows: for

$$\varphi(m) = \frac{C}{m^{\alpha}} \tag{5}$$

we have that

$$\sum_{m=1}^{\infty} \varphi(m) = C \sum_{m=1}^{\infty} \frac{1}{m^{\alpha}} = 1{,}937$$

and hence

$$\frac{1}{\zeta(\alpha)} = \frac{1}{\sum_{m=1}^{\infty} \frac{1}{m^{\alpha}}} = \frac{C}{1,937} = \frac{\varphi(1)}{1,937}$$

$$= \frac{1,287}{1,937} = 0.6644$$

(cf. Table 1) corresponding to $\alpha = 2.18$.

From the adopted $|j_1 j_2 ... j_k|$ notation we were also able to derive the distributional form of the function $\psi(n)$ = the number of papers with n countries, i.e. the n different countries occurring in the m author affiliations in an m-authored paper $(m = 1,2,3,...)$. We found the result in Table 2.

Table 2. Number of papers with n countries (for "pedagog*").

n	$\psi(n)$ (empirical)	fraction: $\frac{\psi(n)}{1,937}$
1	1,844	0.9520
2	74	0.0382
3	16	0.0083
4	3	0.0015
5	0	0

That the function $\psi(n)$ is fastly decreasing is clear from Table 2. The number (for a ψ of Lotka type as in (5))

$$\frac{1}{\zeta(\alpha)} = \frac{1,844}{1,937} = 0.9520$$

even falls outside the tables mentioned above in Egghe and Rousseau (1990), Egghe (2005a) but a "quick and dirty" method estimates

$$\frac{\psi(2)}{\psi(1)} = \frac{1}{2^{\alpha}} = \frac{74}{1,844}$$

yielding $\alpha \approx 4.64$, a seldom seen high Lotka exponent. Treating ψ as an exponential decrease, we estimate

$$\psi(n) = Ca^{n-1} = \left(\frac{C}{a}\right)a^{n} \qquad (6)$$

which gives

$$a = \frac{\psi(2)}{\psi(1)} = \frac{74}{1,844} = 0.0401$$

, an extremally low number, expressing again the fast decrease.

The above example makes it clear that, when studying number of authors or countries per paper, it would be interesting to have a dataset at our disposal where papers have, on average, more authors. This is so in the case of experimental exact or medical sciences. Therefore we analyzed the next dataset.

II.2 Second dataset

The second dataset that was constructed was the result of a retrieval process in the same EBSCO's Academic Search Elite with the command "enzym∗" in the period 2005 (up to October). The same data, as for the first dataset, were collected with the same exclusion policy as described above. We now ended up with 3,271 documents and for each paper we again used the $|j_1 j_2 ... j_k|$ notation (cf. (4) above). We now received the following results. The empirical $\varphi(m)$= number of papers with m authors was as in Table 3.

Table 3. Number of papers with m authors (for "enzym∗").

m	$\varphi(m)$(empirical)	fraction: $\frac{\varphi(m)}{3,271}$
1	286	0.0874
2	422	0.1290
3	508	0.1553
4	503	0.1538
5	437	0.1336
6	368	0.1125
7	241	0.0737
8	159	0.0486
9	128	0.0391
10	75	0.0229
11	42	0.0128
12	37	0.0113
13	23	0.0070
14	11	0.0034
15	11	0.0034
16	2	0.0006
17	2	0.0006
18	2	0.0006
19	2	0.0006
20	12	0.0037

Note that the value $\varphi(20)$ actually is 12 and it is not the cumulation for $m \geq 20$ (there were no papers with 21 or more authors). It is not important here but we are not able to explain why $\varphi(16) = \varphi(17) = \varphi(18) = \varphi(19) = 2$ and $\varphi(20) = 12$.

From the adopted $|j_1 j_2 ... j_k|$ notation we were again able to derive the distributional form of the function $\psi(n)$= the number of papers with n countries for this dataset. We found the result in Table 4.

Table 4. Number of papers with n countries (for "enzym∗").

n	$\psi(n)$(empirical)	fraction: $\frac{\psi(n)}{3,271}$
1	2,737	0.8367
2	443	0.1354
3	65	0.0199

4	11	0.0034
5	8	0.0024
6	3	0.0009
7	2	0.0006
8	1	0.0003
9	1	0.0003
10	0	0
–	–	–

Again the fast decrease is apparent (although less extreme as in Table 2). Note, however, that these data now are derived from a non-decreasing Table 3 for $\varphi(m)$. Now we have (for a ψ of Lotka type as in (5)):

$$\frac{1}{\zeta(\alpha)} = \frac{2,737}{3,271} = 0.837$$

The mentioned tables in Egghe and Rousseau (1990) or Egghe (2005a) now yield $\alpha = 3.04$, a high value but, evidently lower than in the first dataset. Treating ψ as an exponential decrease, we estimate

$$\psi(n) = Ca^{n-1} = \left(\frac{C}{a}\right)a^n \qquad (7)$$

which gives

$$a = \frac{\psi(2)}{\psi(1)} = \frac{443}{2,737} = 0.1619$$

expressing a fast decrease (but less fast as in the first dataset).

That $\psi(1)$ is an extremally high value is clear from the data: for all $m = 1,2,3,...,$ we have that almost all papers are co-authored by authors <u>all</u> coming from the same country. We leave

open to prove why $\psi(n)$ is decreasing for all n. Intuition: the probability for a new country entering a paper is small and hence, the higher n = #different countries, the lower $\psi(n)$.

III. Country occurrences in multi-authored papers: empirical results

III.1 First dataset

The 1,937 papers on "pedagog∗" (2004 + 2005 up to October) in Academic Search Elite showed the following country occurrences in m-authored papers.

Table 5. $\#_{j,m}$: number of times a country appears j times in a paper with m authors $(j = 1,...,m;\ m = 1,2,3,4,5)$ - first dataset.

j \ m	1	2	3	4	5
1	1,287	66	32	9	9
2		334	14	7	8
3			132	3	8
4				53	4
5					24

We deleted the cases m such that $\varphi(m) < 30$, hence (see Table 1) $m = 1,2,3,4$ or 5 since, when there are that small number of papers, the distribution of $\#_{j,m}$ is not very informative. Table 5 clearly shows that $\#_{m,m} ? \#_{j,m}$ for all $m \geq 2$ and $j = 1,...,m-1$ and that $\#_{j,m}, j = 1,...,m-1$ is decreasing, i.e. (2) is valid. The raw data were as follows: 1,287 papers were single-authored papers. The 367 papers with two authors yielded 33 papers of the form $|11|$ (i.e. 2 different countries – cf. the notation in the Introduction: see (4)) and 334 papers of the form $|2|$ (i.e. the two authors are from the same country). Hence $\#_{1,2} = 66$ and $\#_{2,2} = 334$. For $m = 3$ we found 6 papers of the form $|111|$, 14 of the form $|21|$ (or $|12|$) and 132 of the form $|3|$, whence

$\#_{1,3} = 3x6 + 14 = 32$, $\#_{2,3} = 14$, $\#_{3,3} = 132$. For $m = 4$ we have 3 papers of the form $|211|$ (or permutations of these numbers), 2 papers of the form $|22|$, 3 papers of the form $|31|$ (or $|13|$; from no on we drop the mentioning that all permutations are included) and 53 papers of the form $|4|$, whence the results in Table 5. Finally, for $m = 5$, we have $|221|$ once, $|32|$ six times, twice $|311|$, 4 times $|41|$ and 24 times $|5|$ resulting in the $m = 5$ column in Table 5.

III.2 Second dataset

For the second dataset (enzym* (2005 up to October)) (3,271 papers) we had the following paper configuration (we present it in the form of a Table to save space)

Table 6. Paper configurations for the second dataset: the number between brackets shows the number of papers.

m	configurations (incl. permutations)
1	|1| (286)
2	|2| (405), |11| (17)

3	|3| (466), |21| (39), |111| (3)
4	|4| (445), |31| (41), |22| (13), |211| (4)
5	|5| (356), |41| (37), |32| (37), |311| (4), |211| (3)
6	|6| (280), |51| (43), |42| (22), |411| (1), |33| (12), |321| (9), |3111| (1)
7	|7| (182), |61| (21), |52| (16), |511| (3), |43| (16), |421| (3)
8	|8| (99), |71| (18), |62| (16), |611| (7), |53| (8), |521| (3), |44| (3), |431| (1), |4211| (2), |3311| (1), |3221| (1)
9	|9| (81), |81| (14), |72| (6), |711| (3), |63| (4), |621| (2), |54| (8), |531| (2), |522| (1), |441| (1), |432| (2), |411111| (1), |33111| (2), |32211| (1)
10	|10| (50), |91| (6), |82| (5), |73| (4), |721| (1), |7111| (1), |64| (2), |622| (1), |55| (1), |541| (2), |532| (1), |33211| (1)
11	|11| (25), |101| (4), |92| (2), |83| (1), |821| (2), |8111| (1), |74| (1), |731| (1), |6221| (1), |62111| (1), |5111111| (1), |4421| (1), |4211111| (1)
12	|12| (23), |111| (6), |102| (2), |921| (1), |84| (2), |75| (1), |732| (1), |642| (1)

Again we deleted the cases m such that $\varphi(m) < 30$, hence (see Table 2) $m = 1,2,...,12$ for the same reason as mentioned above (in fact, in Table 6, even for $m = 12$, one can see that many configurations are missing due to the fact that we have very few papers (here 37) and increasingly (with m) many possible configurations.

The data of Table 6 lead to the following Table 7 on the values $\#_{j,m}$ ($j = 1,...,m$; $m = 1,2,...,12$).

Table 7. Values of $\#_{j,m}$ ($j = 1,...,m$; $m = 1,...,12$) – second dataset.

m / j	1	2	3	4	5	6	7	8	9	10	11	12
1	286	34	48	49	48	57	30	43	38	14	26	7

2		405	39	30	43	31	19	23	14	10	9	5
3			466	41	41	34	16	12	13	7	2	1
4				445	37	23	19	9	13	4	4	3
5					356	43	19	11	11	5	1	1
6						280	21	23	6	3	2	1
7							182	18	9	6	2	2
8								99	14	5	4	2
9									81	6	2	1
10										50	4	2
11											25	6
12												23

Now it is clear from Table 7 that $\#_{m,m} ? \#_{j,m}$ for all $j = 1,...,m-1$, $m = 1,2,...,12$ and that $\#_{1,m}$ is the second largest value, i.e. the largest in $\{\#_{1,m},...,\#_{m-1,m}\}$, i.e. (3) is valid. We have the impression that, for higher values of j, up to $m-1$, the values are again increasing. An intuitive explanation is that collaboration of authors of the same country is very likely and hence that, besides $\#_{m,m}$, the numbers $\#_{m-1,m}$, $\#_{m-2,m}$ are relatively high, hence also forcing $\#_{1,m}$, $\#_{2,m}$ to be relatively high. In short, we expect that, in cases where we have enough papers with a large number m of authors, the distribution of $\#_{j,m}$ is J-shaped but we cannot give a rationale for it.

We underline the fact that the $\psi(n)$-type data of the previous section and the $\#_{j,m}$-data of this section are, to the best of our knowledge, new types of informetric data and hence, in the line of the arguments given in Egghe (2005b), have a universal interest.

IV. Country occurrences in multi-authored papers: combinatorial results

It is clear that a complete explanation of the regularities of the distribution of the numbers $\#_{j,m}, j = 1,...,m$ is difficult to give. In this section we will "accept" the fact that

$$\#_{m,m} \; ? \; \#_{j,m} \tag{8}$$

for all $j = 1,...,m-1$ and all m based on the natural assumption that collaboration between authors of the same country has the highest probability.

Neglecting (8) for a while we will prove, in a combinatorial way, that (9) is valid:

$$\#_{1,m} > \#_{2,m} > \#_{3,m} > ... > \#_{m-1,m} = \#_{m,m} = 1 \tag{9}$$

for all m, provided that all $|j_1 j_2 ... j_k|$ situations (cf. Section II) are equally possible and that each possibility occurs only once. This is certainly not true in practise: the situations with k high (and a fortiori the case $|\underbrace{11...1}_{m \text{ times}}|$) has a smaller probability to occur than cases where there are only a few countries involved (e.g. $|(m-i)i|$ where only two countries are involved $(i > 1)$). Nevertheless, an exact proof of (9) under this assumption will shed some light on the experimental results obtained in the previous section.

Theorem IV.1:
For every $m \in \mathbb{N}$, $m \geq 2$ and under the assumption that all cases $|j_1 j_2 ... j_k|$, such that

$$\sum_{i=1}^{k} j_i = m, \tag{4}$$

are equally possible and occur only once, we have that

$$\#_{1,m} > \#_{2,m} > \#_{3,m} > ... > \#_{m-1,m} = \#_{m,m} = 1 \tag{9}$$

is valid.

Proof:

The proof is given by complete induction. For $m = 2$ the result is trivial since in (9) we have $\#_{1,2} = 2\#_{2,2}$, $\#_{2,2} = 1$ (cases |11| and |2|). For $m = 3$ we have, besides |3| the cases |21| and |111|, so $\#_{1,3} = 4\#_{2,3}$ and $\#_{2,3} = \#_{3,3} = 1$ proving (9). Now we could start the induction argument but, for an insight in this induction argument, we will present some more cases of low m-values. For $m = 4$ we have, besides |4|: |31|, |22|, |211| and |1111|. Hence $\#_{1,4} = 7\#_{3,4}$ and $\#_{2,4} = 3\#_{3,4}$, and $\#_{3,4} = \#_{4,4} = 1$ proving (9). For $m = 5$ we have, besides |5|: |41|, |32|, |311|, |221|, |2111|, |11111|. So $\#_{1,5} = 12\#_{4,5}$, $\#_{2,5} = 4\#_{4,5}$ and $\#_{3,5} = 2\#_{4,5}$ and $\#_{4,5} = \#_{5,5} = 1$ again proving (9). Last example: if $m = 6$ we have, besides |6|: |51|, |42|, |411|, |33|, |321|, |3111|, |222|, |2211|, |21111|, |111111|, hence $\#_{1,6} = 19\#_{5,6}$, $\#_{2,6} = 8\#_{5,6}$, $\#_{3,6} = 4\#_{5,6}$, $\#_{4,6} = 2\#_{5,6}$ and $\#_{5,6} = \#_{6,6} = 1$ hence again proving (9). Note that all these cases $|j_1 j_2 ... j_k|$ comprise all permutations of the $j_1, ..., j_k$. Let us now start the induction argument.

Let us suppose that (9) is valid for all numbers of authors per paper which are smaller than or equal to $m-1$ (as we verified already for all $m-1 \leq 6$). We have to prove (9) for the case m, i.e. where the papers have m authors. In the m-case, if we delete in all possible cases <u>one</u> 1 then we, obviously, have all possible situations in the $(m-1)$-case. In total we hence deleted $\#_{m-1}$ 1s where $\#_{m-1}$ denotes the number of $(m-1)$-cases. Hence we have

$$\#_{1,m} = \#_{1,m-1} + \#_{m-1} \tag{10}$$

In the m-case, if we delete in all possible cases <u>one</u> 2 then we, obviously, have all possible situations in the $(m-2)$-case. In total we hence deleted $\#_{m-2}$ 2s. Hence

$$\#_{2,m} = \#_{2,m-2} + \#_{m-2} \tag{11}$$

For all $j \leq \left[\frac{m}{2}\right]$ (where $[x]$ denotes the largest entire number smaller than or equal to x), in the m-case, if we delete in all possible cases <u>one</u> j then we obviously have all possible situations in the $(m-j)$-case. In total we hence deleted $\#_{m-j}$ js. Hence

$$\#_{j,m} = \#_{j,m-j} + \#_{m-j} \tag{12}$$

since $j \le m-j$. For $j > \left\lfloor \frac{m}{2} \right\rfloor$ we hence have $j \ge \left\lfloor \frac{m}{2} \right\rfloor + 1$, hence $j > m-j$ no matter if m is odd or even. Now (12) is still valid but remark that from now on $\#_{j,m-j} = 0$ and hence $\#_{j,m} = \#_{m-j}$. Hence (12) is valid for all $j = 1,2,...,m-1$. Note that, always $\#_{m-1,m} = \#_{m,m} = 1$. We also have, trivially,

$$\#_m > \#_{m-1} > ... > \#_{m-j} \qquad (13)$$

for all $j = 1,2,...,m-1$. We now have, for all $j = 2,...,m-1$, by (12)

$$\#_{j-1,m} = \#_{j-1,m-(j-1)} + \#_{m-(j-1)}$$

$$> \#_{j,m-(j-1)} + \#_{m-j}$$

(induction hypothesis for $m-(j-1) \le m-1$ and (13))

$$\ge \#_{j,m-j} + \#_{m-j}$$

since it is obvious that $\#_{j,m-(j-1)} \ge \#_{j,m-j}$

$$= \#_{j,m}$$

by (12). This ends the induction proof. ,

V. Conclusions and open problems

When looking at country occurrences in multi-authored papers we found the following regularities based on two datasets. The number of papers $\psi(n)$ with n different countries is a fastly decreasing function, independent of the fact that $\varphi(m)$, the number of papers with m authors is decreasing or not.

Let $\#_{j,m}$ denote the number of times a country appears j times in an m-authored paper $(j = 1,...,m;\ \ m = 1,2,...)$. We always have that

$$\#_{m,m} ?\ \ \#_{j,m} \tag{14}$$

for all $j = 1,...,m-1$, i.e. most papers are co-authored by authors from the same country. For bibliographies where there is not much collaboration we have that, for all $m = 2,3...$

$$\#_{1,m} > \#_{2,m} > ... > \#_{m-1,m} \tag{15}$$

while for bibliographies where there is much collaboration we have the weaker relation:

$$\#_{1,m} > \#_{j,m} \tag{16}$$

for all $j = 2,...,m-1$ in general.

An explanation of (15) (hence also of the weaker (16)) is given by a combinatorial argument supposing that all cases $|j_1 j_2 ... j_k|$ are equally possible (i.e. country i occurs j_i times in this paper, $i = 1,...,k$). This, however overestimates the occurrence of many different countries and underestimates the occurrence of few countries. We therefore state as an open problem to prove (15) or (16) in both types of bibliographies, using realistic probabilities for the occurrence of each $|j_1 j_2 ... j_k|$ in each case. Furthermore we consider it a challenge to find a model for $\#_{j,m}, j = 1,...,m$ which we conjecture, for large values of m, to be J-shaped.

We also lack a model for the fastly decreasing function $\psi(n)$= the number of papers with n different countries appearing in the author affiliations and we leave it as an open problem to explain the decrease, even in cases where $\varphi(m)$= the number of papers with m authors is not decreasing.

We believe data on country occurrence were never collected before. The results could be applied to estimate the difference between total and fractional counting of country scores as suggested in Gauffriau and Larsen (2005) but for this we are in need of a realistic model for the function $\#_{j,m}, j = 1,...,m$, for every m.

References

L. Egghe (1993). Consequences of Lotka's law in the case of fractional counting of authorship and of first author counts. Mathematical and Computer Modelling 18(9), 63-77.

L. Egghe (2005a). Power Laws in the Information Production Process: Lotkaian Informetrics. Elsevier, Oxford, UK.

L. Egghe (2005b). Expansion of the field of informetrics: origins and consequences. Information Processing and Management 41, 1311-1316.

L. Egghe and I.K. Ravichandra Rao (2002). Duality revisited: construction of fractional frequency distributions based on two dual Lotka laws. Journal of the American Society for Information Science and Technology 53(10), 789-801.

L. Egghe and R. Rousseau (1990). Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science. Elsevier, Amsterdam, the Netherlands.

L. Egghe, R. Rousseau and G. van Hooydonk (2000). Methods for accrediting publications to authors or countries: consequences for evaluation studies. Journal of the American Society for Information Science 51(2), 145-157.

M. Gauffriau and P.O. Larsen (2005). Counting methods are decisive for rankings based on publication and citation studies. Scientometrics 64(1), 85-93.

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

E. Kranakis and E. Kranakis (1988). Comparing two weighting methods in citation analysis. Unpublished paper. Amsterdam, the Netherlands.

D. Lindsey (1980). Production and citation measures in the sociology of science: the problem of multiple authorship. Social Studies of Science 10, 145-162.

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