MODELLING AND MEASURING MULTILATERAL CO-AUTHORSHIP IN INTERNATIONAL SCIENTIFIC COLLABORATION. PART I. DEVELOPMENT OF A NEW MODEL USING A SERIES EXPANSION APPROACH

C. DE LANGE,* W. GLÄNZEL**

* Catholic University Nijmegen, Education and Research Policy Unit, P.O.Box 9102, NL-6500 HC Nijmegen (The Netherlands)
** Research Association for Science Communication and Information e.V. (RASCI), Johannes-Kepler-Weg 5, D-15236 Frankfurt (Oder) (Germany)

(Received August 30, 1997)

International co-operation has strongly intensified during the last decades owing to rapid developments in scientific communication. Economic, political, and intra-scientific factors also strongly influence international collaboration links among individual countries. Obviously research results of international scientific co-operation are reflected in the documented scientific communication as international co-authorship links in scientific publications. Most bibliometric studies on this issue pertain to the share of international co-authored papers in national publication output and their impact on national and international research, or to the analysis and mapping of the structure of collaboration links. The present study attempts to develop a model to measure and analyse the extent of multilateral international co-authorship links. A new indicator, the Multilateral Collaboration Index (\( p \)) is introduced and analysed as a function of the share of internationally co-authored papers (\( f \)). Based on \( f \) a series expansion approach is applied that can be considered an extension of a fractionation model by Nederhof and Moed and allows classifying the extent of multilateral links both among science fields and among individual countries. The paper is concluded by a first attempt to estimate the errors involved in our approach.

Introduction

During the last decades international collaboration in science has been rapidly growing. Though bibliographic data from the Science Citation Index® of the Institute for Scientific Information (ISI, Philadelphia, PA, USA) reflect a certain 'saturation' in...
international scientific collaboration since 1994, the phenomenon has been the object of many recent studies which pertain to its various economic, political and many intra-scientific causes but which also attempt to create or apply suitable models to measure tendencies in and the impact of internationalising scientific research (e.g., deBeaver and Rosen, 1978, 1979a, 1979b, Schubert and Braun, 1990, Narin and Whitlow, 1990, Lee Pao, 1992, Glänzel and Winterhager, 1992, Luukkonen et al, 1992, Luukkonen et al, 1993).

Research results of international scientific collaboration are usually published in the scientific literature. From the bibliometric viewpoint, this form of documented science communication makes international collaboration measurable. Most bibliometric studies on this issue are thus concerned with the share of internationally co-authored papers in the national publication output and their impact on national and international research, or with the analysis and mapping of the structure of collaboration links. The share of 'international papers' and the analysis of selected collaboration clusters, however, reflect only certain aspects of scientific collaboration.

Among the attempts to quantify multilateral research and to derive rules from defined indicators we mention the fractionation approach by Nederhof and Moed (1993). This approach, originally elaborated for the national level of aggregation, results in a very simple model that describes the fractionation degree, i.e., the weights of countries involved in multinational publications, as a linear function of the number of international co-authored papers. Though Nederhof and Moed have developed this method to estimate fractionation based on on-line retrieved shares on internationally co-authored papers, and although they have also conducted error calculations for their model, it should be stressed that the deviation of empirical data from the regression line should not necessarily be interpreted in terms of perhaps insufficiently good fits or limited validity of the underlying model. Deviations may also express significantly differing collaboration patterns in several countries. Fractional counts are, however, not only influenced by the share of multinational (international) papers, but also by the extent of multilateral international collaboration in the individual countries. In the present study the authors, therefore, attempt to measure and analyse the extent of multilateral international co-authorship links and to compare the results with two particular models based on the countries' share of international publications; this can be considered a methodological extension of the aforementioned fractionation approach by Nederhof and Moed.
Constructing the model

Determining internationally co-authored papers

In the past decades, as the possibilities of computer manipulation of large data sets have developed rapidly, bibliographic databases have become the primary object of research in scientometrics. In this section we describe an indirect method to obtain the share of papers which are the product of international co-operation. The method can be used for the Science Citation Index®, as well as field-oriented, abstract-type databases like Physics Abstracts®. In contrast to a CD-ROM version it is generally not possible for on-line versions of bibliographic databases to obtain directly the number of internationally co-authored papers for a certain (sub)set of publications. We describe this indirect method below, even though it is rather well-known, because we need this description to develop a new parameter that can be used to study bibliometrically the extent of multilaterality in international collaboration.

First one selects a set of publications (e.g., for a certain subfield and for a certain period of time). Secondly for this particular set one determines the nationalities of all authors of these papers (first authors, as well as co-authors). As an example let us assume that we find USA, UKD, FRA, GER, JPN, …, and HUN. As the next step one determines for this publication set the number of publications, which are written by authors of different countries. If our country of interest is Hungary than we obtain

\[ N_{\text{HUN,int}} = N_{\text{HUN,USA}} \cup N_{\text{HUN,UKD}} \cup N_{\text{HUN,FRA}} \cup N_{\text{HUN,GER}} \cup N_{\text{HUN,JPN}} \cup \ldots \]  (1)

where, e.g., \( N_{\text{HUN,USA}} \) denotes the set of papers with authors both from Hungary as well as the USA, and \( \cup \) denotes the union of two sets.

Now we have determined \( N_{\text{HUN,USA}} \) etc., we will introduce the number of international collaboration links \( M_{\text{HUN,int}} \) as the sum of these numbers

\[ M_{\text{HUN,int}} = N_{\text{HUN,USA}} + N_{\text{HUN,UKD}} + N_{\text{HUN,FRA}} + N_{\text{HUN,GER}} + N_{\text{HUN,JPN}} + \ldots \]  (2)

It is important to stress at this point that it is possible in a straightforward manner to obtain \( N_{\text{int}} \) and \( M_{\text{int}} \) empirically for any set of publications under consideration, as outlined above! As we will see in the following sections, it is possible to classify the type of multilateral collaboration, once both parameters have been determined.
Structure of multilateral collaboration

Let us now turn to the underlying structure of these parameters. We introduce the following notations. For convenience sake we drop the prefix for the country under study in the following. Let \( n_i \) then denote the number of a country's publications with co-authors from \( i \) different countries. Thus \( n_1 \) denotes the number of domestic papers, \( n_2 \) the number of bilateral, \( n_3 \) the number of trilateral papers, and so on. Consequently, we have

\[
N = n_1 + n_2 + n_3 + n_4 + \ldots + n_i + \ldots \equiv n_1 + N_{\text{int}}
\] (3)

where \( N \) is the total number of publications for this country, and

\[
N_{\text{int}} = n_2 + n_3 + n_4 + \ldots + n_i + \ldots
\] (4)

By definition the share of internationally co-authored papers \( f \) is the ratio \( N_{\text{int}}/N \). Therefore

\[
N_{\text{int}} \equiv fN = f(n_1 + N_{\text{int}}) = n_1 f/(1-f)
\] (5)

The number of international collaboration links \( M_{\text{int}} \), as defined in Eq. (2), is equal to the number of multilateral co-authorship links determined by country pairs. When for a given country all publications in a certain field are produced in collaboration with other countries \( a, b, c, d, \ldots \), the total number of papers in combination with each of these countries is

\[
m_a = n_a + n_{ab} + n_{ac} + n_{ad} + \ldots + n_{abc} + n_{abd} + n_{bcd} + \ldots + n_{abcd} + \ldots
\]

\[
m_b = n_b + n_{ab} + n_{bc} + n_{bd} + \ldots + n_{abc} + n_{bcd} + n_{abd} + \ldots + n_{abcd} + \ldots
\]

\[
m_c = n_c + n_{ac} + n_{bc} + n_{cd} + \ldots + n_{abc} + n_{acd} + n_{bcd} + \ldots + n_{abcd} + \ldots
\]

\[
m_d = n_d + n_{ad} + n_{bd} + n_{cd} + \ldots + n_{abd} + n_{bcd} + n_{acd} + \ldots + n_{abcd} + \ldots
\]

\[
\ldots
\] (6)

Here \( n_a \) is the number of bilateral papers with authors from the country under study together with authors from country \( a \) only; \( n_{ab} \) is the number of trilateral papers with together with authors from countries \( a \) and \( b \) only; etc. If one looks at Eq. (6), it is easily seen that the number of papers produced by each country pair appears twice; that each trilateral combination appears thrice, etc. Hence we obtain

\[
M_{\text{int}} = m_a + m_b + m_c + m_d + \ldots = n_2 + 2n_3 + 3n_4 + 4n_5 + \ldots
\] (7)
Now we introduce a new variable, the fractional count $N_f$, for which the weights of countries involved in multinational publications are a linear function of the number of countries involved, i.e.

$$N_f = n_1 + n_2/2 + n_3/3 + n_4/4 + \ldots + n_i/i + \ldots .$$

$N_f$ is often used (cf., e.g., Nederhof and Moed, 1993) to avoid double counting of internationally co-authored papers (i.e. if one sums the fractional counts of every country that has published in a certain field and a certain period of time, the result is equal to the world total of publications in that field for that period).

**Series expansion approximation**

In reality the world has a finite number of countries, but in order to develop a model for the publications numbers $N_{int}$, $M_{int}$ and $N_f$ let us assume that this number is infinite. Hence we can write for Eq. (4):

$$N_{int} = \sum_{i=2}^{\infty} n_i = n_1 \sum_{i=2}^{\infty} n_i / n_1 .$$

We can rewrite Eq. (3) and make use of the well-known geometric series expansion of $1/(1-f)$:

$$N_{int} = n_1 f/(1-f) = n_1 (f + f^2 + f^3 + f^4 + \ldots + f^i + \ldots) .$$

Note that the series expansion is allowed, since $f$ ranges between 0 (solely domestic publications) and 1 (entirely international papers). If we compare the right-hand side of Eqs (9) and (10), we see that in an ideal situation the number of papers with international co-authors decreases monotonously with a factor $f$ like a geometric series, i.e.

$$n_{i+1} = fn_i \text{ for all } i \geq 1 .$$

We should stress that this only holds when the number of international papers $N_{int}$ is uniquely determined by $f$, which actually is not the case. However, when one looks at an arbitrary set of publications, in general for large enough samples of publications, it is to be expected that the number of bilateral papers will be a factor larger than the number of trilateral papers, and so on, *ad infinitum*. In such a situation the geometric series expansion might yield a rather accurate approximation, provided there is still a considerable amount of publications with authors from more than, say, 6 or more
countries, because otherwise the contribution higher order terms in the series expansion will be exaggerated.

Substituting Eq. (11) in Eq. (7) we obtain:

\[ M_{int} = \sum_{i=2}^{\infty} (i-1) \cdot n_i = n_1 \sum_{i=2}^{\infty} (i-1) \cdot n_i / n_1 = n_1 \sum_{i=2}^{\infty} i \cdot f^i = n_1 f(1-f)^{-2}. \] (12)

Likewise we get for Eq. (8):

\[ N_f = 1 + f/2 + f^2/3 + f^3/4 + \ldots = - (n_1/f) \ln (1-f). \] (13)

**Linear approximation**

Neglecting the fraction of papers with co-authors from four or more countries (linear approximation), Eq. (13) can be written as:

\[ N_f \equiv n_1 \left( 1 + f/2 + f^2/3 \right). \] (14)

Since \( n_1 = (1 - f) N \) (cf. Eq. (5)) we obtain (neglecting terms of the order \( f^3 \)):

\[ N_f \equiv (1 - f/2 - f^2/6) N = (1 - 0.5 f - 0.167 f^2) N. \] (15)

Whereas Nederhof and Moed (1993) find by method of linear regression through a set of empirical samples

\[ N_f \equiv (1 - 0.5022 f - 0.147 f^2) N \] (16)

which shows that their linear regression results correlate excellently with our linear approximation. It should be noted that Nederhof and Moed used only a few samples, where there was strong multilateral collaboration (Earth & Space Science, and Astronomy). It is doubtful that the linear approximation will hold for fields, like, e.g., experimental nuclear physics, where the multilateral collaboration is dominant due to the fact that most of the research is performed at large international facilities, like CERN. We will come back to the range of values for which Eqs (15) and (16) give a good approximation in a later section after we have introduced the Multilateral Collaboration Index \( \rho \).
Multilateral Collaboration Index

Since the number of collaboration links can not be less than the number of international papers, that is, \( M_{int} \geq N_{int} \) we define the Multilateral Collaboration Index \((\rho)\) as

\[
\rho \equiv (M_{int} - N_{int}) / N_{int}.
\]  
(17)

In verbal terms, \( \rho \) expresses the relative multilateral character in the international collaboration links of a country under study.

Substituting the series expansion approximation of Eq. (12) in the definition of \( \rho \), we arrive at

\[
\rho^{(\infty)} = f/(1-f).
\]  
(18)

We will call this representation of \( \rho \) its full series expansion.

If we assume that the fraction of papers with co-authors from four or more countries can be neglected compared with the share of domestic, bi- and trilateral collaboration then we can omit all expressions of higher order than 1 in the second line of Eq. (5). We therefore call the resulting representation of \( \rho \) its linear or first order approximation. In particular, we have

\[
\rho^{(1)} \equiv (1 + 2f) / (1 + f) - 1 = f / (1 + f).
\]  
(19)

Note that \( \rho^{(1)} \leq 0.5 \) whereas \( \rho^{(\infty)} \) tends to infinity with growing \( f \). The empirical \( \rho \) values are, of course, expected to reflect field-specific peculiarities. Thus multilateral research, e.g., in fields like nuclear science, general and internal medicine — where research is often performed in large international research groups — is much more intense than, e.g., in mathematics, engineering, or chemistry. Both the full series expansion \( \rho^{(\infty)} \) and the linear approximation \( \rho^{(1)} \) are deterministic functions with one free parameter each. Since these functions depend only on the share of international papers; they are not suited to reflect field-specific peculiarities in the extent of multilateral research collaboration. Therefore \( \rho^{(\infty)} \) and \( \rho^{(1)} \) are expected to measure both the countries' and the fields' deviations from an ideal situation. In particular, we can interpret observations as follows. First, note that \( \rho^{(\infty)} > \rho^{(1)} \) for every \( f > 0 \). Thus we can say that when a country has an observed \( \rho \) value lower than \( \rho^{(1)} \), the multilateral collaboration is mainly bi- and trilateral. On the other hand, when \( \rho \) is greater than \( \rho^{(\infty)} \) then multilateral collaboration really dominates, even over bilateral collaboration. This applies both to fields and to individual countries within selected fields. In nuclear
physics, for example, we will find almost all countries above the $\rho^{(\infty)}$ curve, whereas in mathematics most countries will be located below the $\rho^{(1)}$ curve.

Three basic types of multilateral collaboration patterns

The multilateral collaboration index ($\rho$) is expected to be strongly field-dependent. We introduce three basic types of ($f$, $\rho$) charts. These types can be used to classify field-specific collaboration patterns. Type I is the standard type where all points are almost symmetrically scattered with respect to the $\rho=f$ line. Two subtypes may be distinguished. The first one represents the type where the chart is clearly subdivided into three zones with a similar number of elements. The second subtype, where all elements are located between the two curves, would correspond to the ideal case as assumed in the previous section.

Type II reflects the situation where bi- and trilateral collaboration dominates. Again, the first subtype corresponds to the case where the area above the $\rho^{(\infty)}$ curve is (almost) empty, the second one to the case, where (almost) all countries are located below the $\rho^{(1)}$ curve.
Type III (again with two subtypes) can be defined in an analogous manner for dominating multilateral collaboration. This classification scheme is sketched in Fig. 1. Of course, the borderlines between these types and subtypes are ‘fuzzy’. Thus any classification of science areas or subfields based on the above types and subtypes is somewhat arbitrary, but, even such ‘rough’ classification schemes may increase understanding of the particular collaboration patterns in individual science areas and their change in time.

**Relation between the fractional count \( N_f \) and the multilateral collaboration index \( p \)**

In the full series expansion we find from Eqs (10) and (11)

\[ M_{\text{int}} = N_{\text{int}}/(1-f). \]  

(20)

Substituting \( 1-f = N_{\text{int}}/M_{\text{int}} \) in the expression for \( N_f \) in Eq. (13) we get

\[ N_f = -\left( n_1/f \right) \ln \left( N_{\text{int}}/M_{\text{int}} \right) = \left\{ N M_{\text{int}}/(M_{\text{int}}-N_{\text{int}}) \right\} \ln \left( M_{\text{int}}/N_{\text{int}} \right) \]  

(21)

since \( n_1 = (1-f)N \) and \( f = (M_{\text{int}}-N_{\text{int}})/M_{\text{int}} \). By definition \( p \equiv (M_{\text{int}}-N_{\text{int}})/N_{\text{int}} \) (cf. Eq. 17), so we get \( M_{\text{int}}/N_{\text{int}} = p + 1 \). Now we can make use of the well-known series expansion

\[ \ln x = (x-1)-(x-1)^2/2 + (x-1)^3/3-(x-1)^4/4 + (x-1)^5/5-... \]  

(22)

for \( 0 < x \leq 2 \), to rewrite Eq. (21) as

\[ N_f = (N/p)(p^2/2 + p^3/3-...) = N(1-p/2 + p^2/3-...) \]  

(23)

Thus we find a rather elegant expression that relates the fractional count \( N_f \) and the multilateral collaboration index \( p \). Now we can compare the various approximations for \( N_f \) of Eqs (13), (15), (16) and (23) in combination with the empirical values of \( N_f \) found by Nederhof and Moed (1993). In Table 1 their samples are given.

In Figure 2 different approximations of \( N_f \) are plotted as a function of \( f \) together with the empirical values given in Table 1. The four different curves in Figure 2 represent (i) the full series expansion of Eq. (13) [C], (ii) the linear approximation of Eq. (15) [A], (iii) the full series expansion of Eq. (23) [D], and (iv) the linear approximation of Eq. (23) [B] omitting terms higher than \( p^3 \). In addition, the empirical data by Nederhof and Moed [E] are presented in Figure 2, but the linear regression approximation by Nederhof and Moed in Eq. (16) has not been plotted because it is indistinguishable from Eq. (15).
Table 1
Empirical data relating the fractional count ($N_f$) with the share of internationally co-authored papers ($f$) for a number of fields and countries (from Nederhof and Moed, 1993)

<table>
<thead>
<tr>
<th>Country</th>
<th>Field</th>
<th>$f$</th>
<th>$N_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>Education</td>
<td>0.022</td>
<td>0.989</td>
</tr>
<tr>
<td>USA</td>
<td>Psychology</td>
<td>0.053</td>
<td>0.972</td>
</tr>
<tr>
<td>DEU</td>
<td>Psychology</td>
<td>0.088</td>
<td>0.955</td>
</tr>
<tr>
<td>JAP</td>
<td>Biology</td>
<td>0.101</td>
<td>0.948</td>
</tr>
<tr>
<td>NLD</td>
<td>Psychology</td>
<td>0.169</td>
<td>0.911</td>
</tr>
<tr>
<td>USA</td>
<td>Earth &amp; Space</td>
<td>0.196</td>
<td>0.895</td>
</tr>
<tr>
<td>NLD</td>
<td>Engineering</td>
<td>0.351</td>
<td>0.811</td>
</tr>
<tr>
<td>NLD</td>
<td>Earth &amp; Space</td>
<td>0.470</td>
<td>0.736</td>
</tr>
<tr>
<td>SWE</td>
<td>Astronomy</td>
<td>0.599</td>
<td>0.637</td>
</tr>
</tbody>
</table>

We see that both the linear and the full series expansions essentially give the same value for $N_f$ in the range $0 < f < 0.5$ and $0 < p < 1$. It is rather surprising that we obtain the same value for $N_f$ for even for larger values of $f$ and $p$ for all approximations (due to the fact that the series expansions converges rapidly in this range of $f$ and $p$). The full series expansion expression for $N_f$ given in Eq. (23) as a function of $p$ starts to deviate considerably for $p > 0.6$. If we do not break off the series expansion after the first few terms. Therefore it can be concluded that for the fractional count $N_f$ the series expansion in $f$ leads to an extremely rapidly converging approximation, even for large $f$ and $p$. It is remarkable that both the linear approximation and the full series expansion give essentially the same fit with the empirical data of Nederhof and Moed. On the other hand, however, the series expansion for $N_f$ as a function of $p$ does not result in an accurate approximation for larger values of $f$ and $p$.

Concluding remarks

The introduced model proved appropriate to reflect both the particular collaboration patterns in individual science areas and their change in time. However, we have also to look at error estimates. Conclusions drawn from the above results can only be valid if the random error is within reasonable limits. Especially the standard error of the Multilateral Collaboration Index can be problematic. In particular, we have $D(f) \approx \{(1-f)/N\}^{-\frac{1}{2}} \lesssim N^{-\frac{1}{2}}$. The standard error of the share of international papers of a
Fig. 2. Relation between fractional count ($N_f$), Share of international papers ($f$) and Multilateral Collaboration Index ($p$), based on different approximations, and on the equation $p^{(\infty)} = f(1-f)$

(A: $N_f = 1-f/2-f^2/6$, B: $N_f = 1-p/2+p^2/3$, C: $N_f = \text{func}(f)$, D: $N_f = \text{func}(p)$, E: empirical data for $N_f$ according to Nederhof and Moed)

country with, for example, 50 (100) publications is thus less than 0.07 (0.05). The calculation of the standard error of the Multilateral Collaboration Index is more difficult since $M_{int}/N_{int} - 1$ is only an asymptotically unbiased estimate of the theoretical $p$ value. However, if we proceed from the assumption that $p$ is approximately an ideal variable, that is, if we assume the collaboration links to have a geometric distribution with parameter ($q = n_1/N = 1-f$) over international papers, then we obtain $D(p) = p^{(\infty)}N_{int}^{1/2}$ for the standard error of the Multilateral Collaboration Index, provided $N_{int} > 0$. In verbal terms, the standard errors of the two indicators are of the same order if both $N_{int}$ and $n_1$ are not to small. These estimates can be used for the statistical reliability tests in cross-national analyses within selected fields.

* 

This study was in part supported by an Alexander von Humboldt research fellowship. The co-author wishes to acknowledge the support. The authors thank E. J. Rinia for stimulating discussions.
References


NARIN, F., WHITLOW, E. S. (1990), Measurement of scientific cooperation and co-authorship in CEC-related areas of science. Commission of the European Communities. Belgium.


