GEOGRAPHICAL PROXIMITY AND SCIENTIFIC COLLABORATION

J. S. KATZ

Science Policy Research Unit, University of Sussex, Falmer, Brighton (UK) BN1 9RF

(Received April 26, 1993)

Geography, economic, socio-political and language are considered to be factors that effect the level of research collaboration. However, to-date no technique has been developed to isolate the effect of geographical proximity from the other factors. This paper presents a methodology for specifically examining geographical effects on intra-national scientific collaboration. An investigation of intra-national university-university collaboration in Canada, Australia and the United Kingdom using this technique demonstrates that research cooperation decreases exponentially with the distance separating the collaborative partners.

Introduction

In recent years, the scientific community and science policy analysts have become more interested in scientific research collaborations. Numerous initiatives have been aimed at increasing collaboration among individual scientists, as well as, improving the links between science and technology by fostering research collaboration across sectors – in particular, between universities and companies. Furthermore, most governments have been keen to encourage their scientific community to increase their levels of participation in international collaboration. However, few, if any, policies directed towards increasing collaboration are designed to take account of any effect of geographical separation on the level of collaboration. This may be due to the fact that there is little more than suggestive evidence to show that geographical proximity has a significant influence on scientific cooperation. This paper presents evidence, derived using a novel methodology, that within a country's scientific community the number of collaborations decreases rapidly as a function of the distance separating research partners.

Before we direct our attention to the effect of geography, let us briefly review what previous observers of collaboration have identified as the main factors encouraging scientific cooperation. The literature on this topic is extensive and it suggests that the factors promoting collaboration include: (1) changing patterns of
funding, scientific popularity, visibility and recognition, (3) rationalization of scientific manpower, (4) the demands of complex large scale instrumentation, increasing specialization in science, the degree of advancement of a particular discipline, (7) the professionalization of science, (8) the need to gain experience, train researchers and sponsor protégés, (9) the desire to increase cross-fertilization of ideas and techniques, and (10) a decrease in spatial distance. This list is far from complete. Scientific collaboration is a social process and probably there are as many reasons for researchers to collaborate as there are reasons for people to communicate.

Most investigators agree that scientific collaboration tends to begin informally and is often the result of causal conversation. This communication may lead to successively greater commitments to cooperate in a manner similar to the role that communication plays in courtship. Using various analytical techniques, researchers have shown that geographical, social and political factors seem to lead to more collaboration since it is likely to produce more informal communication. Many of these investigators have drawn the conclusion that the closer two potential collaborators are, the more likely they are to initiate informal communication. However, verification of this claim requires a methodology that specifically isolates the geographical effects from other factors. No such a methodology appears to have been reported previously.

Techniques commonly used to map collaboration

A number of methods have been used to map the intensity of scientific research collaboration between both individuals and countries. Usually this analysis is based on bibliometric data which is normalized to remove the effect of such factors as country or institutional size. Finally, the data are processed using a clustering process to generate the collaboration maps. Many normalization procedures use non-linear weighting techniques favouring values that are either high or low on the spectrum of values. Frequently, multidimensional scaling is used as the clustering technique to produce collaboration 'maps' which are subjectively analyzed to see if geographical, political, economical or cultural factors seem to be influencing the collaborative patterns. Let us examine the individual steps in the process in more detail.

Peters and Van Raan explored the effect of different normalization techniques on two-dimensional maps produced by carrying out a multidimensional scaling of co-
words derived from chemical engineering papers. They used three techniques: (1) the Jaccard index, (2) the proximity index and (3) the inclusion index. With these, they produced three maps. Each map was composed of the same words but each map had very different spatial characteristics. The maps were presented to experts in chemical engineering to see if they could determine which one provided the most realistic representation of activity in the field. There was no consensus among the experts. Peters and Van Raan concluded that it was unclear which index should be used. One difficulty with this approach which does not seem to have been investigated concerns the reliability of expert 'mental' maps, that is the mental images that experts have of the relationships that exist between entities. There is no evidence to suggest that even a knowledgeable observer is capable of mentally storing and reconstructing the complex relationships that exist in maps derived empirically from a large number of variables.23

Let us examine multidimensional scaling a little more closely. In addition to exploring collaborative activity by investigating co-authorship linkages, multidimensional scaling has been used for a variety of other purposes in bibliometric studies including: (1) displaying relationships between publishing and citing patterns,24 (2) exploring the social structures of scientific specialties,25 (3) examining author co-nomination,26 (4) investigating journal similarities,27 and (5) mapping co-word relationships.22, 28 In essence, multidimensional scaling allows the investigator to explore the underlying structure of similarities or dissimilarities between entities in a co-occurrence relationships usually expressed in the form of a square matrix. The relationships in an $n \times n$ matrix exist in multidimensional space and multidimensional scaling techniques project these relationships into a lower dimensional space (typically one, two or three dimensions) using linear regression techniques.29

Generally speaking, in collaboration studies those collaborators who exhibit a high co-occurrence in multidimensional space are placed in close vicinity to one another in a lower dimensional space. However, the projection is seldom, if ever, perfect and frequently significant distortion occurs because of the information loss that is incurred during the mathematical compression of co-occurrence relationships from many dimensions to a few dimensions. This can result in collaborators who are not situated near one another in multidimensional space being placed near one another in low dimensional space.

The potential for producing non-linear effects when normalizing data and the possibility of introducing proximity distortions using multidimensional scaling to
produce collaborative maps raises at least two important questions. Which of the variety of normalization procedures is the most appropriate for examining collaboration? And how many of the influencing factors that have been claimed to be revealed by collaboration maps are merely artifacts during the normalization multidimensional scaling process? These and other concerns point to the need for an alternative approach which would allow the effect of geographical proximity on scientific collaboration to be investigated in isolation from other factors.

A non-distorting method for analyzing the effect of geographical proximity

The method to be described for measuring the effect of geographical proximity on scientific collaboration was devised as part of a research project that used bibliometric techniques to assess *intranational* (i.e. within a single country) university-university collaboration within the United Kingdom, Canada and Australia.\(^{30}\) The general approach involved an analysis of corporate addresses from articles, notes and reviews\(^ {31}\) recorded in the *Science Citation Index* and published by UK and Australian universities between 1981 and 1990 and by Canadian universities between 1984 and 1990.\(^ {32}\) A university-university collaboration was defined 'as any publication in which two or more domestic university names appeared in the corporate address field'.

Generally speaking the process for examining the relationship between geographical locations and the amount of collaboration involved three steps. First, all the university addresses were unified to eliminate data-entry errors and to assign standardized names to individual universities. The unified data were used to produce a publication matrix (publication × university) that was algebraically manipulated to make a co-occurrence collaboration matrix (university × university). The next step involved producing a radial distance matrix for each country that specified the geographical distance between each pair of universities in the country. Finally, the number of collaborations that occurred in each of a series of distance intervals was summed and plotted to determine if there was a relationship between geographic proximity and the frequency of collaborations. The following section outlines the rational for using a pair-wise collaboration counting method and details the technique used to produce the collaboration and radial distance matrices.
Counting collaborations

There is no well-formulated method for counting institutional collaborations. The fractional counting method used in co-authorship analysis was examined for its suitability for counting institutional collaboration. Advocates of this method fractionate \( \frac{1}{n} \) each publication and assign each on \( n \) authors listed on a paper \( \frac{1}{n} \) of the paper. There is a problem using this techniques for counting institutional collaboration. To properly allocate the contribution of each author to the appropriate institution, one has to be able to determine which author is associated with which institution. This is not possible with data from the Science Citation Index since it does not provide information on individual author-institution affiliations. Also, in cases where the number of authors is greater than the number of institutions, there is no way of determining which author resides at which institution and therefore how many authors are located at each institution.

Consideration was given to the possibility of fractionating each paper according to the number of institutions involved. However, this seemed to be unfair in certain instances. Consider the example of a publication with three authors and two institutions where two authors reside at one institution and one author at the other institution. If each author participated fully in this collaboration, then it would seem unreasonable to allocate the institution with two authors \( \frac{1}{2} \) of the credit when its researchers actually contributed \( \frac{2}{3} \) of the effort.

After due consideration, it was decided that the counting technique should indicate how 'attractive' an university appears to be to the rest of the domestic university community for forming collaborative linkages. A technique which seemed to adequately measure the collaborative attractiveness of a university was one which counts the number of 'pairings' a university has with other universities. Consequently, the technique which was adopted was to count the number of two-way collaborations that occur in each publication. For example, a paper listing four universities A, B, C and D would be counted as having six two-way collaborations: A-B, A-C, A-D, B-C, B-D and C-D. In general, if the number of institutions is \( n \), the number of two-way collaborations, \( c \), is given by

\[
c = \frac{n(n-1)}{2}
\]
In general, this value is not very large as illustrated by the fact that an analysis of the UK, Canadian and Australian publications showed that over 93 per cent of all intranational university-university collaborations involve only two universities.\textsuperscript{30}

A two-way collaboration matrix is derived from a publication matrix. A publication matrix is a binary, university by publication, matrix where the entries in each cell are either a 1 or 0. Each row in the publication matrix represents one publication; each column represents one university. If a given university participates in a publication its column entry for that publication is assigned a 1, otherwise it is assigned a 0. Formally, a publication matrix can be defined as follows:

Given a set of universities, \( U = (u_1, u_2, u_3 \ldots u_n) \), and a set of publications, \( P = (p_1, p_2, p_3 \ldots p_m) \), the publication matrix \( Q \) is a \( n \times m \) binary matrix where \( Q_{ij} = 1 \) if \( u_i \) is listed in \( p_j \), otherwise \( Q_{ij} = 0 \).

By way of illustration, consider an example consisting of five publications from a community of four universities: A, B, C, and D. Assume the publications involved the following university collaborations: (1) AB, (2) AD, (3) BD, (4) ACD and (5) BCD. Using the above definition of a publication matrix, the result would be

\[
Q = \begin{pmatrix}
1 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 \\
1 & 0 & 1 & 1 \\
0 & 1 & 1 & 1 \\
\end{pmatrix}
\]

where the columns indicate university A, B, C and D, respectively, and the rows indicate paper one to five, respectively.

A collaboration matrix is a square, university by university, matrix that records the number of two-way collaborations that one university has with another university. This matrix is derived from the publication matrix by multiplying the transpose of the publication matrix, \( Q^T \), by the original matrix, \( Q \). Therefore, the collaboration matrix, \( C \), is given by

\[
C = Q^T Q
\]

The diagonal cells, \( C_{ii} \) where \( i = j \), in the resultant publication matrix contain the number of papers in which \( u_i \) participated; the remaining cells, \( C_{ij} \) where \( i \neq j \),
contain the number of two-way collaborations between \( u_i \) and \( u_j \). If we use the example publication matrix above, the resultant collaboration matrix would be

\[
C = \begin{pmatrix}
1 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 \\
0 & 1 & 1 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 \\
1 & 0 & 1 & 1
\end{pmatrix}
= \begin{pmatrix}
3 & 1 & 1 & 2 \\
1 & 3 & 1 & 2 \\
1 & 1 & 2 & 2 \\
2 & 2 & 2 & 4
\end{pmatrix}
\]

In the resultant matrix, \( C \), both the rows and columns indicate universities A, B, C, and D, respectively. The diagonal elements show that A was involved in three multi-university publications, B in three, C in two and D in four. The contents of the first row indicate that A was involved in one two-way collaboration with B, one with C and two with D.

**Measuring the distance between collaborators**

The distance between collaborators was tabulated in a radial distance matrix. The radial distance matrix, \( R_i \), is a square, university by university, matrix that contains the distance, measured in kilometres (or miles), between pairs of universities within a country. The diagonal cells have been set to zero; \( R_{ij} = 0 \) where \( i = j \), i.e. the distance between any university and itself is zero. The remaining cells, \( R_{ij} \) where \( i \neq j \), were assigned the distance in kilometres between \( u_i \) and \( u_j \). Universities in the same city were arbitrarily assigned a distance of 25 km.

The distance between universities were determined by using an accurate scale map of the country and measuring the distance between cities in which the universities are located. In order to facilitate country comparisons, the inter-university distances for each country were normalized by dividing by each cell in the matrix by the distance separating the two most distant universities in that country.

The normalized radial distance matrix is used in conjunction with the collaboration matrix to determine the effect of distance on collaboration. For our purposes this was done by tabulating the number of two-way collaborations that occurred in each 0.1 distance interval (0.00–0.09, 0.10–0.19, ... 0.90–1.00) between 0.0, the minimum, and 1.0, the maximum. This is most easily accomplished if the ordering of the names of the institutions are the same in both matrices. Then each \( R_{ij} \), were \( i \neq j \), determines which of ten interval counters \( C_{ij} \) is to be added. Finally,
the totals in the interval counters are used to make an X–Y (distance-collaboration) scatter plot.

**Effect of geographical proximity on intranational university-university collaboration**

The number of two-way collaborations in the UK and Australia between 1981 and 1990 and in Canada between 1984 and 1990 is plotted against the normalized geographical distance separating the universities in each country (see Fig. 1). For convenience a fixed distance interval of 0.1 or 10 per cent of the country size was chosen and the number of collaborations that occurred within each interval was counted. We define country size to be the furthest distance between any pair of universities. This is approximately 700 km in the UK, 2100 km in Australia and 4800 km in Canada; therefore each 0.1 normalized interval represents 70 km, 210 km and 480 km, respectively. An exponential regression analysis of collaborations versus distance was performed and in each instance there was a good fit (t-test < 0.01).

![Graph](image)

**Fig. 1.** Effect of geographic proximity on intranational university-university collaboration (U = United Kingdom, C = Canada, A = Australia). One distance interval equals 10 per cent of the maximum distance separating the two most distant universities.

Table 1 gives the values of the y-intercept and exponent for the exponential regression line for each country that best fit the equation, \( y = ae^{bd} \) where \( y \) is the number of two-way collaborations, \( a \) is the y-intercept, \( b \) is the exponent and \( d \) is the normalized distance.
Table 1
Comparative values for \( y = ae^{bd} \)

<table>
<thead>
<tr>
<th>Country</th>
<th>( a \times 10^3 )</th>
<th>( b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>18</td>
<td>(-5.0 \pm 0.8)</td>
</tr>
<tr>
<td>Canada</td>
<td>24</td>
<td>(-3.8 \pm 0.5)</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>133</td>
<td>(-5.2 \pm 0.6)</td>
</tr>
</tbody>
</table>

The exponents for the UK and Canada are very similar while the exponent for Australia is approximately 20 per cent lower. This suggests that the effect of geographical proximity in Australia, although still significant, is somewhat less because the exponential decrease with respect to distance is not as steep. The reason for this difference is primarily due to fact that the geographical distribution of the universities in Australia is more dispersed. Australia has a very large land mass – approximately seventy-seven per cent that of Canada and thirty times that of the UK – and approximately half as many universities as Canada. Over seventy-five per cent of the Australian universities are situated in the south eastern part of the country while three universities are located in the south west (all in Perth), one in the north and none in the interior. The effect of the geographical distribution of the universities can be seen partly by the fact that there are no collaborations data points in 0.60 – 0.69 interval.

In the UK it can be seen that the actual number of collaborations in the 0.0 and 0.1 interval is less than might be expected. This can be explained by two facts. First, the radius of greater London is approximately 40 km (25 mi) or approximately one distance interval in diameter and the University of London is involved in a large number of collaborations. Second for the purpose of our study, all collaborations with the individual colleges of the University of London were counted as collaborations with the University of London. In fact, the individual colleges are essentially autonomous and the number of collaborations that occurred with each one could have been counted separately which would have increased the number of two-way collaborations in the first interval by a significant amount.

This graph provides convincing evidence that university-university collaboration decreases exponentially with distance and therefore occurs more frequently with partners who are geographically closer than with those further away. In general, more than 60 per cent of all two-way university-university collaborations in Canada
and Australia and more than 40 per cent in the UK occur within a collaboration radius equal to less than 20 per cent of the country size. Another way of looking at this is to note that the number of collaborations in Canada and the UK decreases approximately by one-half each time the collaboration radius is increased by an interval equal to 15 per cent of the country size.

This analysis was performed on all the publications recorded in the Science Citation Index for a ten-year period in the United Kingdom and Australia and a six-year period in Canada across all scientific fields. If an analysis were to be performed on individual science fields, we might well find that the result is affected by the geographical location of centres of excellence. For example, the effect of geography on collaboration in physics might be modulated by the geographical location of universities that have exceptional experimental facilities such as high-energy particle accelerators or collaborations in clinical medicine may be modulated by universities with experts who are proficient in new surgical procedures or who possess complex instrumentation such as CT scanners and nuclear magnetic imaging devices.

Conclusions

A methodology has been presented that uniquely separates the effect of geographical proximity on intra-national university-university collaboration from other factors (e.g. social, political and linguistic factors). The frequency of research collaboration between domestic universities in the United Kingdom, Canada and Australia decreases exponentially with the distance separating the research partners. This provides strong evidence to support the notion that informal, 'face-to-face' communication may be an essential ingredient in research collaborations and that factors such as greater geographical distance with the additional travel cost and time involved are impediments to collaboration.

One possible improvement that could be made to this technique is to replace the radial distance matrix with a cost-time-distance matrix that is, a matrix composed of information derived from the cost per unit time per unit distance travelled between collaborators. This might provide some insight into the effect of cost and travel time on collaboration. However, since the cost and time are dependent on both the mode and class (e.g. economy, first class rail) of travel, a number of arbitrary decisions would be required.

This technique can also be used to examine the effect of geography on inter-institutional collaboration between and within other sectors. A research project is
under way to investigate this effect in the UK with a particular emphasis on collaborations between institutions in the public sector and companies. Finally, it might be possible to evaluate international collaboration using this methodology. However, it is likely that the number of geographical locations may be very large and the complexity of the problem might have to reduced by examining collaborative activity only between those institutions situated in major metropolitan cities.

In conclusion, geographical proximity plays a significant role in collaboration. Designers of policies that encourage scientific cooperation must take this factor into account if they wish to promote collaboration over a large geographical area. Finally, the insights presented in this article must be viewed in context because Griffith and Miller\(^\text{39}\) have noted the paradoxical phenomenon that 'individual scientists may be reluctant at one extreme, to travel seventy-five feet to utilise another person's store of knowledge but, at the other extreme, would willingly travel hundreds or thousands of miles to communicate with other persons in other circumstances'.

* 

The author is grateful to Ben Martin, Francis Narin, Diana Hicks and Leon Katz for helpful comments on early drafts of the paper. He also acknowledges the support of the Economic and Social Research Council (ESRC), the Advisory Board for the Research Councils (ABRC) who funded the bibliometric database used in this study. Finally, the author extends his gratitude to Ravi Maithel and the Saskatchewan Research Council for supporting his mid-life pursuit of a doctoral degree.

References and notes


21. For example, consider the normalization technique used by Schubert and Braun (1990) in their analysis of international collaboration. They used a relative strength measure, $r_{ij}$, to normalise their data. This measure was given by:

$$r_{ij} = \frac{n_{ij}}{\sqrt{n_i n_j}} = \sqrt{\frac{n_{ij}}{n_i}} \frac{n_j}{n_i}$$

where $n_i$ and $n_j$ are the total numbers of papers published by countries $i$ and $j$, respectively, and $n_{ij}$ is the number of papers co-authored by $i$ and $j$. It can be seen that as $n_{ij} \rightarrow n_i$ and $n_j$ and $n_i < n_j$ then

$$r_{ij} \rightarrow \sqrt{\frac{n_{ij}}{n_j}}$$

which implies an non-linear dependence of $r_{ij}$ on $n_{ij}$. A similar argument can be made for the normalization method that Luukkonen, Person and Sivertsen (1990) used in their examination of international collaboration. They normalised their data using observed collaboration frequency versus expected collaboration frequency ratios. This ratio was given by

$$Y_{ij} = \frac{C_{ij} T}{C_i C_j}$$

where $C_{ij}$ is the number of collaborations between countries $i$ and $j$, $C_i$ and $C_j$ are the total number of collaborations each country has with all other countries, and $T$ is the total number of collaborations among all countries.


23. Given a map with $n$ objects (words, authors, countries, etc) there are $n(n-1)/2$ connections between pairs of objects to evaluated. For example, given a map with 10 objects there are 45 possible pairs. For each pair there are two possible decisions to make: (1) should a pair of objects be connected i.e. is there a relationship the pair of objects, and (2) if there is a relationship how intense or strong is it? Therefore, the minimum number of decision to be made will be 45 (i.e. there are no connections between any pair of objects). The maximum number of decisions will be $45 \times c$, where $c$ is the maximum number of discrete values the intensity or strength of the connection can assume. If each connection can have 3 values (e.g. high, medium and low) and if all possible pairs are connected, there will be 135 decisions to make. Notice, we have not taken into account the decisions required to determine where each object should be placed on the map. An experiment that you can perform is to choose ten family members. Ask member each to draw a map indicating the strength of the
relationship between each pair in the group. Finally, compare the maps for uniformity. Peters and Van Raan asked observers to examine co-word maps that depicted the relationship between more than 50 words. Is it reasonable to assume that an observer can construct a 'mental map' which requires a minimum of 1225 decisions?

29. Each element in a row (or column) in a co-occurrence matrix is assumed to be independent (or orthogonal) to every other element in that row (or column). Therefore, each element is considered to be a value for a vector which has a dimensionality equal to the number of elements in the row (or column).
31. These publication types contain substantial contributions to scientific knowledge.
32. The UK and Australian data was derived from 1981–1990 *Science Citation Index* tapes purchased by the Science Policy and Research Evaluation group at SPRU and the Research School of Social Sciences at the Australian National University for use in the academic research performance indicators project. Special permission was granted by ISI's UK office to download the Canadian data from the BIDS database at the University of Bath in return for feedback on the performance of their on-line enquiry software which was then under development.
35. We analyzed twenty-two Australian, forty-six Canadian and fifty-two UK universities.
36. The University of London is in greater London and not only is it the largest university in the UK, it also has the largest number of collaborations in the UK. With the exception of a couple of smaller universities like the Universities of Reading and Surrey, most universities are more than 70 km (43 miles) away and their collaborations with the University of London would not be counted in the 0.0 to 0.1 interval.
37. The university of London is composed of a number of autonomous colleges. These are Imperial College, Birkbeck College, University College, Kings College, Queen Mary College, Westfield College, Chelsea College and Royal Holloway and Bedford College.
38. Twenty per cent of the country size is 140 km in UK, 410 km in Australia and 960 km in Canada.