Measuring the relative intensity of collaboration within a network

Reference:
Fuchs Joel Emanuel, Sivertsen Gunnar, Rousseau Ronald.- Measuring the relative intensity of collaboration within a network
Scientometrics: an international journal for all quantitative aspects of the science of science and science policy- ISSN 1588-2861 - 126:10(2021), p. 8673-8682
Full text (Publisher's DOI): https://doi.org/10.1007/S11192-021-04110-X
To cite this reference: https://hdl.handle.net/10067/1816130151162165141

Measuring the relative intensity of collaboration within a network

[1]Joel Emanuel Fuchs, [2]Gunnar Sivertsen & [3,4]Ronald Rousseau

[1]University of Wuppertal, Gaußstr. 20, 42119 Wuppertal, Germany; jfuchs@uni-wuppertal.de
ORCID:0000-0002-9951-3492

Nordic Institute for Studies in Innovation, Research and Education, NO-0608 Oslo, Norway;
gunnar.sivertsen@nifu.no
ORCID: 0000-0003-1020-3189

[3]University of Antwerp, Faculty of Social Sciences, B-2020 Antwerp, Belgium;
ronald.rousseau@uantwerpen.be
ORCID:0000-0002-3252-2538

[4]KU Leuven, Facultair Onderzoekscentrum ECOOM, Naamsestraat 61, Leuven B-3000, Belgium;
ronald.rousseau@kuleuven.be
ORCID:0000-0002-3252-2538

ABSTRACT

In research fields like informetrics or patent research, international comparisons are part of their core business. One class of measures used in such comparisons consists of affinity indices. Here, relations of shares are calculated to express the relation between two actors X and Y. Even though this group of affinity indices already has many members, we think that based on pure logic, they still miss the point. For this reason, we introduce the *RIC* index. This index compares two shares: the share of actor Y within a given set X to the share of Y within the complement of X. After motivating and introducing this new index, we show some of its basic characteristics, the most interesting one being that increased collaboration between two countries leads to an increase in the value of *RIC*. Moreover, this index is asymmetric. The *RIC* index is illustrated with the examples of China and the USA within the global network of collaboration 2000-2020.

Keywords: international collaboration; activity index; probabilistic affinity index; citation index; Relative Intensity of Collaboration; RIC

Introduction

Measuring collaboration, and more generally co-occurrence is a long-standing research topic in informetrics. Sonnenwald (2007) defined research collaboration as “the interaction taking place within a social context among two or more scientists that facilitates the sharing of meaning and completion of tasks with respect to a mutually shared, superordinate goal”. In informetric studies, the term collaboration most often refers to jointly publishing authors, universities, sectors (as in Triple Helix studies), or countries. We will use the term ’actors’ when referring to any of these. In all these cases there is a network, consisting of nodes and links. Mathematical formulae reflecting degrees of collaboration can be used for other networks as well, such as sociograms and friendship networks (Wasserman & Faust, 1994) where they then figure as co-occurrence measures (at least this is our opinion).

Although there is always an underlying network, this can be totally ignored, taken partially into account, or taken as completely as possible into account. When collecting data on the percentage of single-authored papers, for instance, one completely ignores the network structure. In this paper, we take the link structure partially into account, while focusing on two nodes. In network theory one distinguishes between ego networks, starting with or focusing on one particular node, and networks without a focal node. This distinction also applies to collaboration networks.

A widely used indicator of collaboration

Comparisons between equal-sized partners are easy to perform, but comparisons between very different-sized partners are often biased. This happens in most networks where nodes have a limited and different capacity to link to other nodes. For instance, in a network with humans as nodes and links between those who know each other in person as family members, a person can only have a limited number of links. In this article, we have countries or universities as nodes and collaborated articles as links and corresponding weights. These weights will be heavily influenced by the available scientific workforce. For instance, Germany will have more scientists available for collaborations than Belgium, and China will have even more. Hence, comparing the absolute number of collaborations between countries or universities has limited value. Balassa (1965) was one of the first, who in a study on trade between actors, did not compare absolute numbers, but used an index consisting of shares instead.

Frame (1977) introduced this idea in the information sciences as the activity index of country C with respect to a given field F (and with respect to the world, W) as AI(C, F, W) = (country C’s share in the world’s publication output in the given field F)/(country C's share in the world's publication output in all science fields), see also (Rousseau, 2018, 2019, Rousseau & Yang, 2012). Applying this formula to collaboration studies the Activity Index (AI) becomes:

$$AI(X,Y) \coloneqq \left(\frac{C_{XY}}{C_X}\right) / \left(\frac{C_Y}{T}\right) \tag{1}$$

where C_{XY} denotes the number of collaborations between countries X and Y; C_X (C_Y) the total number of collaborations country X (Y) has with all other countries; and T the total number of pairwise country collaborations in the set of publications under study. Hence T is the sum of the weights of all links in the collaboration network, with countries as its nodes. The AI compares Y's share of collaboration with X to Y's share of collaboration in the whole network. Although we are not aware that this index has actually been applied in collaboration studies, we use it as a starting point to introduce other indices.

One of the best-known formulae to measure collaboration was introduced in Luukkonen et al. (1992). It is widely used to the degree that it has gone through obliteration by incorporation. An example of its use without reference to the original publication is the biannual report *Science and Engineering Indicators* published by the National Science Board (2019) in the USA.

To study the relative importance of country X for country Y while taking other relations into account, Luukkonen et al. (1992) applied an asymmetric formula, denoted here as AOER (Asymmetric Observed to Expected Ratio) and defined as:

$$AOER(X,Y) \coloneqq \left(\frac{C_{XY}}{C_X}\right) / \left(\frac{C_Y}{T-C_X}\right) \tag{2}$$

where C_{XY} denotes the number of collaborations between countries X and Y (they used whole counting); C_X the total number of collaborations country X has with all other countries, C_Y the total number of collaborations country Y has with all other countries, and T the total number of pairwise country collaborations in the set of publications under study. Hence T is the number of links in the weighted collaboration network, with countries as its nodes. It is assumed that C_X and C_Y are different from zero and for all countries X, C_{XX} is set equal to zero.

It was shown in Chinchilla-Rodriguez et al. (2021) that formula (2) used in (Luukkonen et al., 1992) is one from a whole family of similar indicators, referred to as probabilistic affinity indices, used to study collaboration. These formulae differ among them on two accounts: whether or not the diagonal values of the collaboration matrix containing the weights of the links are set equal to zero or not, and whether normalization for a country's size is applied. Within each group, different choices can be made about how to take diagonal values into account and about how to normalize. Following Luukkonen et al. (1992) we will do neither.

Schubert and Glänzel (2006) referred to formula (2) as the co-authorship preference of country X towards country Y and denoted it by $QA(X,Y)$.

They explain the exclusion of country X from the world total T as follows: if the self-collaboration of each country, denoted as C_{XX}, would be non-zero, the first share of $QA(X,Y)$ would change to $\frac{C_{XY}}{C_X - C_{XX}}$. By the same scheme, the second share would change from $\frac{C_Y}{T}$ to $\frac{C_Y}{T - C_X}$. So, in the denominators of each share, the sums exclude the case involving X. This brought them to formula (2).

We observe that weighting can be done by whole counting, as in Luukkonen et al. (1992), but the mathematical formula can also be applied in the case of fractionalized counting. In that case, the symbols C_{XY}, C_X, and C_Y denote weights associated with collaborated articles between X and Y, collaborated articles between X and another country (any, not just Y), and collaborated articles between Y and another country (again, any, not just X). The symbol T is then the sum of the weights associated with all links in the network under study.

Furthermore, it was shown in (Rousseau, 2021) that formula (2), does not reflect a relative (and absolute) increase in the numbers of collaborated articles. We think that this is an undesired property for a measure that is designed to measure collaboration within a network. For this reason, we propose another indicator, the Relative Intensity of Collaboration (RIC), that meets this requirement.

RIC - a modification of the AOER-index

To understand the need for modification, consider the following set representation, originally due to Charles Dodgson, better known under his pseudonym, Lewis Carroll, which is for our purposes much clearer than a Venn diagram (Dodgson, 1896; Rousseau, 1998).

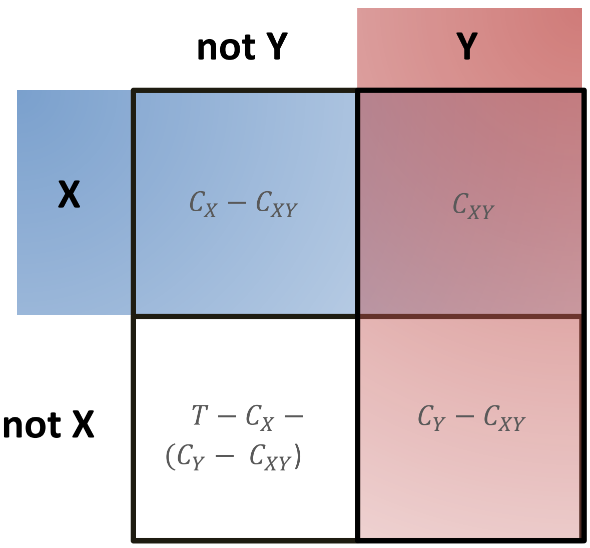

Fig. 1 A Carroll diagram illustrating the collaboration link set

As Carroll diagrams are less known than Venn diagrams, we explain their meaning. The two upper squares represent the set of bilateral links involving country X, the two lower squares represent bilateral links not involving X (the complement of X). The two squares on the right represent the set of bilateral links involving country Y and similarly, those on the left represent links not involving Y (the complement of Y). Hence starting on the top right and moving counter-clockwise the weights associated with each rectangle are: C_{XY}, $C_X – C_{X,Y}$, $T-C_X-C_Y+C_{X,Y}$, and C_Y-C_{XY}. Recall that T is the sum of the weights associated with all links in the network

Formula (2) was said in (Luukkonen et al., 1992) to reflect the relative collaboration importance of country Y to country X, calculated as the share of country Y within all collaborations of country X divided by the share of country Y within all collaborations in the network (minus the collaborations of X). So, the size of the X-Y relation within X's network is compared to the size of Y within the whole network, excluding X. Yet, Figure 1 shows that formula (2) does not actually do so as it still includes a part of X, namely the joint publications of X and Y as part of C_Y.

The definition we propose corrects this. It can be formulated in words as the ratio of the share of the collaborations of A and B within all collaborations of A to the share of collaborations of B within all collaborations of the system excluding collaborations of A. Formulated as a formula, the Relative Intensity of Collaboration (RIC) becomes:

$$RIC(X,Y) := \left(\frac{C_{XY}}{C_X}\right) / \left(\frac{C_Y - C_{XY}}{T - C_X}\right) = \frac{C_{XY} \cdot (T - C_X)}{C_X \cdot (C_Y - C_{XY})} \quad (3)$$

Figure 1 shows that formula (3) corresponds to the ratio of the upper right part divided by the whole upper part, over the lower right part divided by the whole lower part. We will refer to it as the relative intensity of collaboration of actor X to actor Y (in that order), actor X being the focus, denoted as *RIC(X,Y)*. We see that *RIC(X,Y)* ≥ 0, but that there is no theoretical upper limit.

Before studying properties of *RIC* we first consider the exceptional cases. In these cases, *RIC* should not be applied because the relation between X and Y is obvious from the start. The first class of special cases consists of the zero-weighted links: these are the cases, $C_{XY} = 0, C_X = 0, C_Y = 0$ or $T = 0$. From the last to the first: if $T = 0$, then our network would only consist of nodes without a link. We would not have a proper international collaboration network. If $C_X = 0$, then node X does not possess links. X would be isolated and can be excluded from the network. The same holds if $C_Y = 0$. This leaves the case that $C_{XY} = 0$, i.e., there is no link between the nodes X and Y. Then $RIC(X,Y) = \frac{0 \cdot (T - C_X)}{C_X \cdot C_Y}$ = 0. This is a valid result, although, as already stated, obvious from the start. Note that in this last case we have already excluded that C_X or C_Y are equal to zero.

The second class of special cases consists of those to which we refer as monopoly-like links. These are the cases $C_{XY} = T, C_X = T$ and $C_Y = T$ with the condition $T \neq 0$. If $C_{XY} = T$, then the network consists only of two nodes and one weighted link, namely X-Y. We get $C_X = C_Y = C_{XY} = T$, so we would divide through zero if we calculated *RIC(X,Y)*. But this case is again obvious from the start. Similarly, if $C_X = T$, we have an ego network with X as its center. This means, $C_Y = C_{XY}$, as every node only links to X. Again, we would divide through zero. As a result, we exclude this case, too. Finally, we have the case $C_Y = T$. Similarly, as described before, we get $C_X = C_{XY}$. When applying *RIC* to this case, we get $RIC(X,Y) = 1$. So, this case is obvious too but does not have to be excluded.

The last class of special cases we examine consists of the single-linked ones (as we call them). These are the cases $C_X = C_{XY}$ with $C_Y \neq C_{XY}$ and $C_Y = C_{XY}$. In the first case, X is only linked to Y and in the second case, Y is only linked to X. The first case is no problem because although X has only one link (that to Y), we get $RIC(X,Y) = \frac{T-C_{XY}}{C_Y-C_{XY}}$. More problematic is the second case. If Y has only one link to X, it has no links to the complementary set of X. In other words, we divide through zero, if we want to calculate *RIC*. So, we have to exclude this case as well.

So, at last, we obtain this specification of *RIC*:

$$RIC(X,Y) \coloneqq \frac{C_{XY} \cdot (T - C_X)}{C_X \cdot (C_Y - C_{XY})} \text{ if } T \geq C_X > 0, \;\&\; T \geq C_Y > C_{XY}$$

The indicator *RIC* can be written as a function of formula (2), the co-authorship preference of country X towards country Y, as:

$$RIC(X,Y) = \frac{\left(\frac{C_{XY}}{C_X}\right)}{\left(\frac{C_Y - C_{XY}}{T - C_X}\right)} = \left(\frac{C_Y}{C_Y - C_{XY}}\right) \cdot AOER(X,Y)$$

Similar to formula (2), also formula (3) is an asymmetric formula in the sense that its value for X and Y (in this order) differs from its value between Y and X (in that order), as the focal set differs.

We next come to the main reason to introduce *RIC(X,Y)*. In these calculations, we will use the following symbols: $a = C_{XY}, x = C_X - C_{XY}, y = C_Y - C_{XY}$ and $t = T - C_X - C_Y + C_{XY}$. Using this notation we have:

$$RIC(X,Y) = \frac{a \cdot (t+y)}{(x+a) \cdot y} \text{ and similarly, } RIC(Y,X) = \frac{a \cdot (t+x)}{(y+a) \cdot x}$$

showing clearly the asymmetry of *RIC*.

The indicator *RIC* leads to a measure that is sensitive for an absolute and a relative collaboration increase between country X and country Y.

From the formula $RIC(X,Y) = \frac{a \cdot (t+y)}{(x+a) \cdot y}$ we immediately have the following four results, by keeping three variables constant and varying the remaining one.

1) If the collaboration weight of the link X-Y, i.e., the number *a*, increases and all other link weights stay constant, then *RIC(X,Y)* strictly increases, because the ratio *a/(a+x)* increases, with limit 1, when *a* increases. This is the result announced in the title of this section.

2) If the collaboration weight of X with other countries than Y increases, i.e., *x* increases and all other link weights stay constant, then *RIC(X,Y)* strictly decreases.

3) If the collaboration weight of Y with other countries than X increases, i.e., *y* increases and all other link weights stay constant, then *RIC(X,Y)* strictly decreases.

4) If the collaboration weight of collaborations between countries other than X of Y, i.e., the number *t*, increases, then *RIC(X,Y)* strictly increases.

Note that we studied the case of an absolute increase, but as all other data stayed invariant, this also meant a relative increase.

From the first result, we conclude that $RIC(X,Y)$ is a monotone increasing function of $a = C_{XY}$. It is strictly increasing, if $x > 0$. This is the case, as long as X has links to other nodes than Y.

Table 1. An example showing the influence of increased collaboration between countries X and Y on *RIC(X,Y);* the symbol n denotes an absolute increase in the value of C_{XY}.

n	C_{XY}	C_X	C_Y	T	$RIC_n(X,Y)$
0	1	3	4	100	11.111
1	2	4	5	101	16.667
2	3	5	6	102	20.000
3	4	6	7	103	22.222
4	5	7	8	104	23.810
5	6	8	9	105	25.000

6	7	9	10	106	25.926
7	8	10	11	107	26.667
10	11	13	14	110	28.205
100	101	103	104	200	32.686
500	501	502	504	600	33.201

As $\lim_{n\to\infty} RIC_n(X,Y) = \frac{y+t}{y}$, in this example *RIC*$_n$*(X,Y)* tends to 97/3 = 33.233 .

Other properties of *RIC(X,Y)*

In this section, we prove some other properties enjoyed by *RIC*. Notations are used as introduced in the previous section. Special cases as discussed above are excluded.

Proposition 1. If X increases its collaborations with others by an amount of *p* > 0 at the cost of collaborations between others (hence *x* increases by an amount *p* and *t* decreases by *p*), and also *a* ≠ 0 stays the same (hence *y* stays the same, too), then *RIC(X,Y)* decreases.

Proof. We have to check if $\frac{a\cdot(t+y)}{(x+a)\cdot y} > \frac{a.(t-p+y)}{(x+p+a).y}$.

This inequality is equivalent with $(t+y)\cdot(x+p+a) > (x+a)\cdot(t+y-p)$ or $(t+x+y+a)\cdot p > 0$, which is clearly correct.

The following proposition 2 does not hold for formula (2) as shown in (Rousseau, 2021), but it does hold for the intensity indicator *RIC*. This is another example of the “good” behavior of *RIC*.

Proposition 2

Assume that country Y increases its links with other countries (not X) by an amount $p, a > p > 0$, at the cost of links with country X, so *x* and *t* stay the same. Then *RIC(X,Y)* decreases, unless *x* = 0, in which case *RIC(X,Y)* was 1 and stays 1.

Proof. We first consider the case $t = 0$. Then *RIC* is $\frac{a}{x+a}$ and becomes $\frac{a-p}{x+a-p}$, which is obviously smaller if $x \neq 0$. If x and t are both zero then *RIC* is 1 and stays 1. We now consider the general case ($t > 0$).

We have to show that $\frac{a\cdot(t+y)}{(x+a)\cdot y} > \frac{(a-p)\cdot(t+y+p)}{(x+a-p)\cdot(y+p)}$. This is equivalent with:

$$a \cdot (x+a-p) \cdot (y+p) \cdot (t+y) > y \cdot (x+a) \cdot (a-p) \cdot (t+y+p)$$

or

$$axt + a^2t + yxt + y^2x + ypx > apt$$

As $a > p$, we know that $a^2t > apt$. Hence, this proves Proposition 2.

Finally, we consider a special case.

Proposition 3. A balanced case

If there are n countries, and each country collaborates exactly once with each other country in a 2-country publication, and there are no other types of publications, then for every X, Y, *RIC(X,Y)* = 0.5.

Proof. $C_{XY} = 1$. As there are n countries $C_X = n-1$, for every country *X*. Now the total number of links in the network is $\binom{n}{2} = \frac{n(n-1)}{2}$. Then:

$$RIC(X,Y) = \frac{\left(\frac{C_{XY}}{C_X}\right)}{\left(\frac{C_Y - C_{XY}}{T - C_X}\right)} = \frac{C_{XY} \cdot (T - C_X)}{C_X \cdot (C_Y - C_{XY})} = \frac{1.\left(\frac{n(n-1)}{2} - (n-1)\right)}{(n-1).(n-2)}$$

Hence:

$$RIC(X,Y) = \frac{\frac{n^2 - 3n + 2}{2}}{n^2 - 3n + 2} = 0.5$$

This proves this proposition.

Remark, as *RIC* is scale-invariant, the above result also holds if each country collaborates exactly *c* times (c being a natural number different from zero) with each other country in 2-country publications.

A comparison based on real-world data

To compare the properties of *RIC* compared to the *AOER* with the use of real-world data, we use a matrix with collaboration frequencies among the 20 largest contributing countries to journals indexed by *Web of Science* in the years 2000-2020. The direct data source is Web of Science as represented in Clarivate's *InCites*. A collaboration is counted each time author addresses in two different countries appear in the same publication. Among the 20 countries, we illustrate the relations between China and the USA and between the two countries and four other major collaborating countries and how the relative intensity of collaboration among them has developed during the period.

The policy interest of our example is related to the growth of China within the global network. Ten years ago, China surpassed the United Kingdom as the USA's largest collaboration partner in science. The USA now has twice as much collaboration with China as with the UK. However, the relative intensity of collaboration between the USA and China has been declining since it peaked in 2016. China is seeking other partners, among them the UK. This recent development is only visible if an indicator of relative intensity of collaboration is used.

Figure 2 shows the *RIC* applied on the relations of the USA to four other major collaborating countries – relative to the activity in the network between 20 large countries. Figure 3 shows the same relations when the *AOER* is applied.

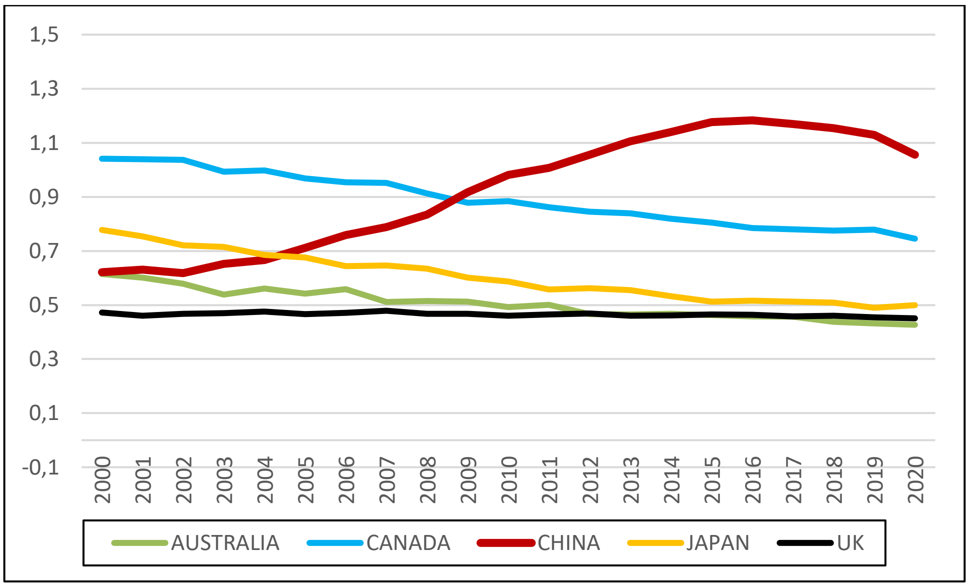

Fig 2 The relative intensity of collaboration (*RIC*) from the perspective of the USA in relation to four major collaborating countries. The USA is X in *RIC(X,Y)*.

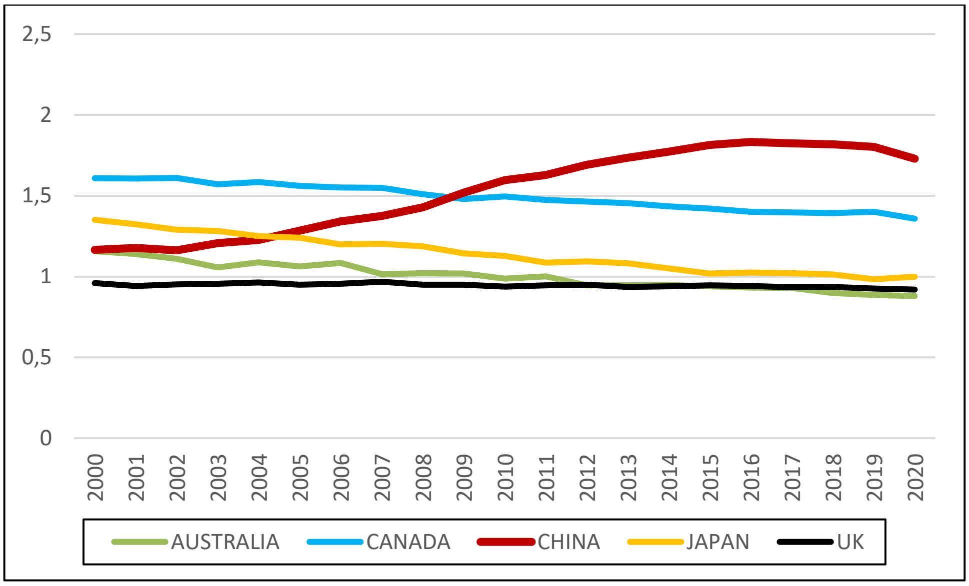

Fig 3 The relative importance (*AOER*) of four major collaborating countries from the perspective of the USA. The USA is X in *AOER(X,Y)*.

Figures 4 and 5 compare the same indicators by applying the perspective of China in the relations to the same five countries.

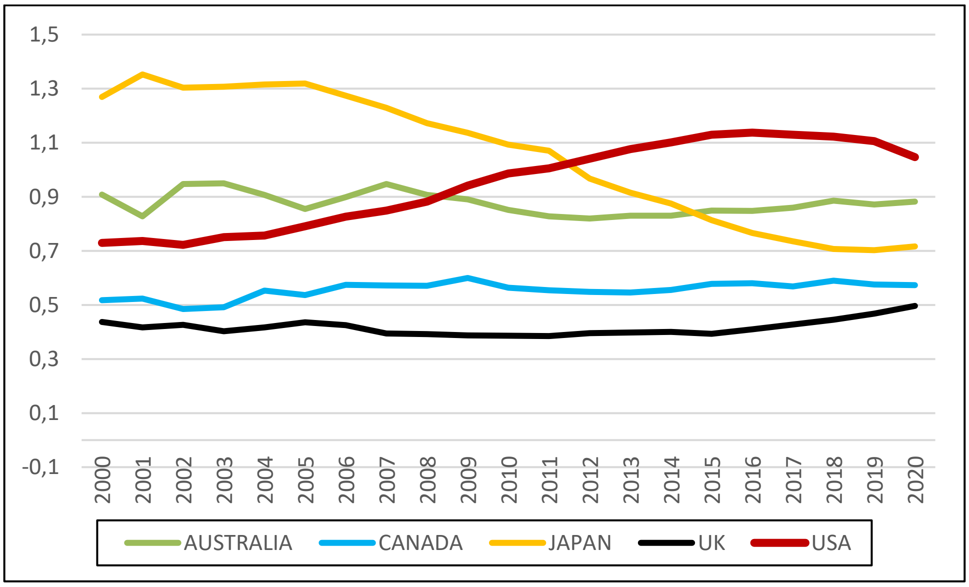

Fig 4 The relative intensity of collaboration (*RIC*) from the perspective of China in relation to five major collaborating countries. China is X in *RIC(X,Y)*.

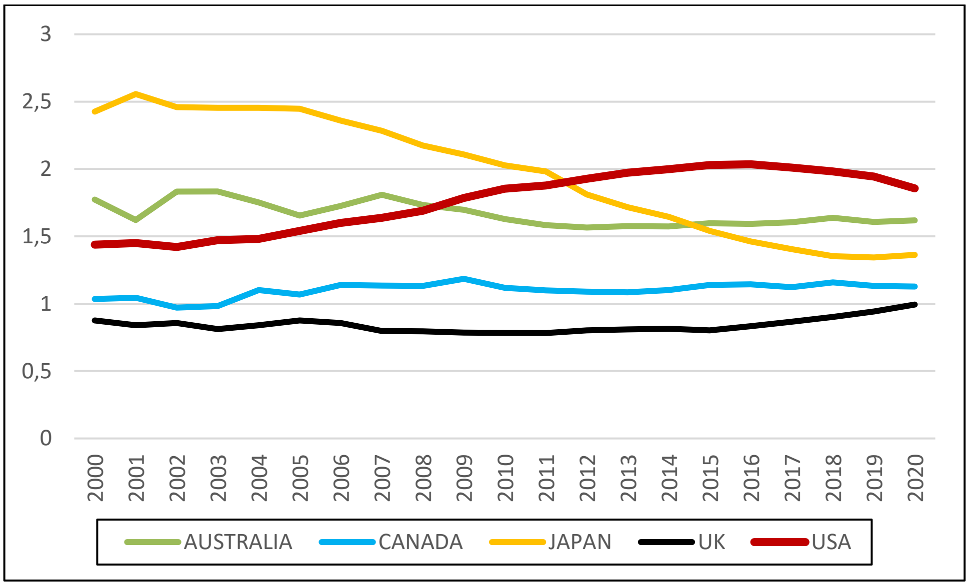

Fig 5 The relative importance (*AOER*) of five major collaborating countries from the perspective of China. China is X in *AOER(X,Y)*.

We observe that the effect of the modification leading from *AOER* to *RIC* is a general decrease in the estimated values. The scales become comparable when applying the *RIC* from the perspective of two different countries. However, the indicator does not show symmetric values in their bilateral relations. The trends and relative positions of the countries remain the same when applying *RIC* instead of *AOER*.

Discussion and conclusion

We recall the two most important properties of *RIC,* namely that it is an asymmetric indicator as in general *RIC(X,Y)* ≠ *RIC(Y,X)*, and that increased collaboration between two countries leads to an increase in the value of *RIC*.

The collaboration intensity measure *RIC* does not take collaborations within actors, such as countries or universities, into account. In theory, this could be done, but then the upper left part of Fig. 1 would include (in the case of universities) collaborations within one university (X) as well as collaborations with other universities (except university Y). This would lead to an indicator that is at best weak, and in fact meaningless in our opinion. This remark leads to the problem to find a meaningful indicator that takes intra-actor collaborations into account. Note that this problem does not arise for author collaborations. Moreover, as all indicators studied here are collaboration indicators this implies that they do not tell anything about the possible domination of a country in a field. Indeed, domination could imply that this country is involved in 'all' collaborations, but it could also mean that this country does not have to collaborate with others.

In (Rousseau, 2021) we asked if one can invent a collaboration measure for which an increased collaboration between two countries leads to an increase in the collaboration measure? Here we provided a positive answer to this question.

We have shown that the *RIC* has many desirable properties which we intend to use in future investigations. In the end, we leave it to the readers to decide which indicator they want to use.

Acknowledgments

The authors thank Lin Zhang and Grant Lewison for their helpful comments and Ying Huang for providing the China-USA dataset. They also acknowledge useful comments from two anonymous reviewers.

REFERENCES

Balassa, B. (1965). Trade liberalization and "revealed" comparative advantage. *The Manchester School of Economic and Social Studies* 32, 99–123.

Carroll, L. (1896). *Symbolic Logic: Part 1: Elementary*. Reprinted by Dover Books (1958) as part of *Symbolic Logic and the Game of Logic.*

Chinchilla-Rodríguez, Z., Bu, Y., Robinson-García, N., & Sugimoto, C. (2021). An empirical review of the different variants of the probabilistic affinity index as applied to scientific collaboration. *Scientometrics*, https://doi.org/10.1007/s11192-020-03815-9

European Patent Office (1.2.2021). Patent statistics, Granted patents per field of technology and per country of residence for each individual year from 2011 until 2020 (XLS). Retrieved from *https://www.epo.org/about-us/annual-reports-statistics/statistics.html*

Frame, J.D. (1977). Mainstream research in Latin America and the Caribbean. *Interciencia*, 2(3), 143-148.

Luukkonen, T., Persson, O., & Sivertsen, G. (1992). Understanding patterns of international scientific collaboration. *Science, Technology, & Human Values*, *17*(1), 101-126.

National Science Board. (2019). *Science & Engineering Indicators Publications. Output: U.S. Trends and International Comparisons*. https://ncses.nsf.gov/pubs/nsb20206/

Rousseau, R. (1998). Venn, Caroll, Karnaugh en Edwards. *Wiskunde & Onderwijs, 24*(95), 233-241.

Rousseau, R. (2018). The F-measure for research priority. *Journal of Data and Information Science, 3*(1), 1-18.

Rousseau, R. (2019), Balassa = revealed comparative advantage = activity. *Scientometrics, 121*(3), 1835 –1836.

Rousseau, R. (2021). Going back in time: Understanding patterns of international scientific collaboration. *Journal of Scientometric Research, 10*(1), 126-129.

Rousseau, R., & Yang, LY. (2012). Reflections on the activity index and related indicators. *Journal of Informetrics*, 6(3), 413- 421.

Schubert, A., & Glänzel, W. (2006). Cross-national preference in co-authorship, references, and citations. *Scientometrics, 69*(2), 409–428.

Sonnenwald, D.H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology, 41*, 643-681.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis.* Cambridge (UK): Cambridge University Press.