# Impact analysis of domestic and international research collaborations: a Malaysian case study

**Hui Xuan Tan · Ephrance Abu Ujum · Kwai Fatt Choong · Kuru Ratnavelu**

**Abstract**  This paper presents a comparative impact analysis on collaborative research in Malaysia. All analyses were conducted using ISI-indexed journal articles published in the 10-year period spanning the years 2000–2009. The publication growth and distribution of domestic versus international Malaysian-addressed collaborative articles was examined. Then, a three-pronged approach was used to compare the research performance between international and domestic research for the top ten high-productivity subject categories. Firstly, the *potentiality* of collaborative research impact is determined using the Mann–Whitney–Wilcoxon and Bootstrap Kolmogorov–Smirnov tests. Then, the Hirsch and Egghe indices were computed for each subject category to estimate the distance needed to bridge the gap between international and domestic research. Lastly, the composition of researchers was measured using the internationality index. We discuss how the findings of our methodology help advise collaborative research strategies that will contribute to better research performance in the leading scientific categories.

H. X. Tan (✉) · E. A. Ujum · K. Ratnavelu
Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia
e-mail: huixuan@um.edu.my

E. A. Ujum
e-mail: ephrance@siswa.um.edu.my

K. Ratnavelu
e-mail: kuru@um.edu.my

K. F. Choong
SEGi University, Kota Damansara, Selangor, Malaysia
e-mail: kwaifatt@yahoo.com

## Introduction

Recent advancements in communication and transportation technologies increases interactions and networking opportunities (Sonnenwald 2007) amongst even the most geographically dispersed organisations, institutions and partners (Castells 1996; Heimeriks and Vasileiadou 2008; Katz and Hicks 1997). This motivates scientific collaboration, an academic process that connects different sets of knowledge and competencies into novel ideas (Heinze and Kuhlmann 2008). According to Sonnenwald (2007), this process can also be defined as an "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".

Since the pioneering works on collaboration in scientific research by Derek J. deSolla Price (Price 1963; Price and Beaver 1966), Henry Small (Small 1973; Small and Griffith 1974), Eugene Garfield (Garfield 1963, 1964), Belver Griffith (Griffith et al. 1971, 1974) and Donald deB. Beaver (Beaver and Rosen 1978, 1979a, b), amongst others, there have been many attempts made to investigate the structure and dynamics of scientific collaboration, the factors motivating scientific collaboration, the distinctions between different types of collaboration, and their effects on scientific productivity and quality.

Scientific collaboration allows the exchange and sharing of access to funding (Beaver 2001; Heffner 1981), resource and equipment (Beaver 2001; Melin 2000; Thorsteinsdóttir 2000) amongst collaborators, which in turn reduces time and cost (Katz and Martin 1997), and increases efficiency and productivity (Beaver 2001; Thorsteinsdóttir 2000), in research. As larger research groups often draw from a larger reservoir of knowledge (Beaver 2001; Thorsteinsdóttir 2000), each member may possess different expertise that may benefit the group (Beaver 2001; Katz and Martin 1997; Melin 2000; Thorsteinsdóttir 2000). Strategically, collaboration is also a means to gain prestige or visibility (Beaver 2001; Crane 1972; Katz and Martin 1997) to compete with leading national and international research rivals in an increasingly competitive arena. Furthermore, large-scale collaborations between countries, regions, institutions or research fields often form to confront important global issues and threats quickly (e.g. to combat the health threat posed by Severe Acute Respiratory Syndrome in 2003), or even to address fundamental scientific questions (e.g. the world's leading particle physics research centre, European Organisation for Nuclear Research, CERN, was one of Europe's first joint endeavours).

Now, more than ever, it is also critical for researchers to constantly publish new findings; it contributes to the stock of knowledge and development of research, as well as aids in a country's bid to compete for global visibility in science. The number of citations and papers published in high-impact journals as indexed in peer-reviewed databases (such as *Thomson Reuters Web of Knowledge* and *Sciverse Scorpus*) are used as performance indicators for the advancement of a researcher's career (Cave 1997; Kaur et al. 2013). More notably, they are used as measures to gauge research performance impact on an institutional level as seen in many recognised global university rankings such as the *QS World University Rankings*, *Shanghai Jiao Tong* and the *Times Higher Education World University Rankings* (Aguillo et al. 2010; Buela-Casal et al. 2007).

These factors triggered a dramatic growth in international collaborations and internationally co-authored papers worldwide (Georghiou 1998; Glänzel 2001; Miquel and Okubo 1994). The same behaviour can be observed across all areas of research (National Science Board 2012; Wagner and Leydesdorff 2005). Also affected are scientific collaborations in certain Asian nations as seen in country-specific studies in regards to international

collaborations (Anuradha and Urs 2007; Basu and Kumar 2000; He 2009; Niu and Qiu 2014; Wang et al. 2013; Zhou and Glänzel 2010).

However, aside from several papers on international research in Malaysia with respect to selective research fields such as Davarpanah (2009) and Low et al. (2014), the growth of international collaboration and their resulting citation impact is virtually uncharted in the context of Malaysian research. The question of how much international and domestic collaborative research in Malaysia differs in terms of impact and potential has rarely been addressed as well. This gives us a unique opportunity to further this line of enquiry for one of the eight emerging Asia-8 economies that is changing the landscape of research and academic publications in Asia (National Science Board 2012).

Malaysia's burgeoning Science and Technology (S&T) and Research and Development (R&D) presence in South East Asia is aided by the rising numbers of scientific publications and collaborations in recent times (National Science Board 2012). Because of this, countries such as Norway are beginning to view Malaysia as a "potentially beneficial collaborator" (Campbell et al. 2014). Given the current climate of research in Malaysia, it will be useful to examine Malaysia's level of scientific productivity and impact to add to existing Malaysian bibliometric studies, particularly those focused on collaboration in research such as Cheng et al. (2013), Kumar and Jan (2013) and Malairaja and Zawdie (2008).

Thus, this paper aims to present a comparative assessment of Malaysian scientific collaborative performance. We have examined the overall growth and distribution of domestic versus international Malaysian-addressed collaborative publications. We have also used a three-pronged methodology for comparative research performance analysis between international and domestic research. This approach assesses the present state and the potential state of collaborative research; and the composition of researchers in specific subject categories (or in the case of this paper, for the top ten high productivity subject categories).

The rest of this paper is structured as follows. Section "Data" describes the source and treatment of the data used in this research. Section "Methods" introduces the techniques and proposed methodological framework used to evaluate the study datasets. This is followed by the presentation of the results and analysis in section "Analysis". In section "Discussion" we discuss the findings and how they help advise collaborative research strategies that will contribute to better research performance or effectiveness in the leading scientific categories. Section "Conclusion" concludes the paper.

## Data

The source data consisted of publication records retrieved from *Thomson ISI Web of Knowledge*'s *Science Citation Index Expanded* (SCI), *Social Sciences Citation Index* (SSCI) and *Arts and Humanities Citation Index* (A&HCI) databases.

A total of 61,513 publication records that were published under a Malaysian address (search criterion Address = "Malaysia") were downloaded for the period 1980–2012 with no limitations regarding attributes besides language (English publications only). The corresponding citation counts reflected *Thomson ISI*'s tally during data retrieval in June 28, 2012.

Additionally, the source data was restricted to research publications of document type 'Article' in order to control for citation rate differences associated to other document types. For example, review articles are generally known to possess higher citation counts and rates compared to articles because reviews generally steer citations away from the articles it discusses (González-Albo and Bordons 2011).

The source data was then parsed to extract all pertinent information. All data processing, calculation, and exploration was conducted via R and Perl scripts.

## Data classification and construction

Each dataset was divided into two samples: one corresponding to articles produced via domestic collaboration (in which all co-authors are affiliated with a Malaysian address), and the other, via international collaboration (in which at least one-coauthor is affiliated with a non-Malaysian address). Effectively, all single-authored articles were excluded from the study. The domestic and international samples were labeled DOM and INT, respectively. The general characteristics and publication growth of the DOM and INT population for all subject categories spanning the years 2000– 2009 are shown in in Table 1 and discussed in subsection "Publication growth and distribution".

As indicated by the mean citations per paper in Table 3, there are diverse citing practices across disciplines. Thus, to provide a fair comparative assessment of DOM and INT citation performance (see subsections "Distribution test results", "Internationality index scores", and "Hirsch and Egghe indices and summarisation of results"), the datasets were constructed according to subject category definitions provided by *Thomson ISI*'s *Journal Citation Reports* (JCR).

Only the top ten subject categories in which both DOM and INT publication volumes and activity are highest and continuant[1] throughout the study period were considered due to insufficient publication data for some of the less productive subject categories. This ensured that the resulting citation distribution did not contain annual gaps that may affect pair-wise analysis between DOM and INT.

Additionally, the selected year range must not be too broad to minimise ageing effects in the synchronous citation count. It was also important to ensure that the end year did not coincide with the most recent years as low levels of citation count for DOM and INT were expected, making it difficult to detect any statistically significant difference between the two samples (if present). Based on these considerations, the study was restricted to a 10-year period spanning the years 2000–2009. The annual breakdown for the top ten subject categories are as shown in Table 2.

## Author classification and disambiguation

Another facet of interest is the configuration of the research workforce, for which a similar dichotomy on the article level is replicated on the author level. For this purpose, author classification was discerned based on available data. Most ISI records after 2006 explicitly state an author's institutional and country affiliation on an individual basis. In such explicit cases, it was straightforward to identify and list Malaysian authors. For the remainder of cases, we implicitly inferred Malaysian nationality of authors from papers that are not affiliated to any other country but Malaysia. The names obtained via explicit and implicit means were classified as *local* in affiliation. All other names not appearing on this list were classified as *foreign* in affiliation.

As part of the data construction process, some rudimentary measures were taken for author disambiguation. There are two possible scenarios in which author keywords present some

---

[1] For which there are no zero publication counts for the entire sequence of years covered by the study period.

**Table 1** Distribution of DOM and INT articles over a decade, 2000–2009

| Parameters | '00 | '01 | '02 | '03 | '04 | '05 | '06 | '07 | '08 | '09 |
|---|---|---|---|---|---|---|---|---|---|---|
| $N_D$ | 384 | 421 | 423 | 457 | 584 | 653 | 764 | 1030 | 1526 | 2289 |
| $N_I$ | 355 | 400 | 430 | 530 | 576 | 689 | 814 | 998 | 1134 | 1653 |
| $C_D$ | 5,066 | 4,940 | 5,398 | 4,831 | 6,135 | 5,480 | 5,933 | 7,217 | 7,064 | 8,843 |
| $C_I$ | 6,580 | 5,513 | 5,631 | 7,465 | 9,567 | 7,378 | 7,979 | 7,787 | 7,949 | 7,808 |
| $\mu(C_D)$ | 13.2 | 11.7 | 12.8 | 10.6 | 10.5 | 8.4 | 7.8 | 7.0 | 4.6 | 3.9 |
| $\mu(C_I)$ | 18.5 | 13.8 | 13.1 | 14.1 | 16.6 | 10.7 | 9.8 | 7.8 | 7.0 | 4.7 |
| $\mu_{1/2}(C_D)$ | 7.5 | 6 | 7 | 6 | 5 | 5 | 4 | 3 | 2 | 2 |
| $\mu_{1/2}(C_I)$ | 8 | 8 | 6 | 6 | 7 | 5 | 4 | 4 | 3 | 2 |
| $A_D$ | 1,084 | 1,148 | 1,259 | 1,478 | 2,593 | 3,010 | 4,036 | 4,897 | 7,717 | 9,217 |
| $A_I$ | 1,327 | 1,427 | 1,435 | 1,953 | 2,464 | 2,462 | 3,368 | 3,886 | 4,717 | 6,537 |
| $\mu(A_D)$ | 3.6 | 3.6 | 3.7 | 3.4 | 3.6 | 3.8 | 3.8 | 3.8 | 3.8 | 3.9 |
| $\mu(A_I)$ | 4.8 | 4.9 | 4.8 | 4.8 | 5.4 | 4.8 | 5.3 | 10.9 | 6.8 | 8.7 |

$N$, $C$, $\mu$, $\mu_{1/2}$, and $A$ denotes the total article count, the total citation count, the mean, the median, and total distinct author keywords, respectively. The subscripts $D$ and $I$ denote the domestic (DOM) and international (INT) samples, respectively

**Table 2** Annual breakdown for Malaysian top ten high-productivity subject categories

| No. | Subject category, $S$ | $N$ | '00 | '01 | '02 | '03 | '04 | '05 | '06 | '07 | '08 | '09 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CRYSTAL | 2326 | 59 | 71 | 115 | 162 | 161 | 263 | 269 | 317 | 340 | 569 |
| 2 | ENGN:ELEC | 971 | 30 | 30 | 47 | 66 | 87 | 92 | 113 | 144 | 157 | 205 |
| 3 | MAT.SCI | 793 | 30 | 38 | 55 | 26 | 56 | 61 | 56 | 126 | 131 | 214 |
| 4 | FOOD.SCI | 791 | 49 | 50 | 49 | 39 | 59 | 58 | 56 | 97 | 146 | 188 |
| 5 | ENGN:CHEM | 727 | 19 | 21 | 19 | 44 | 63 | 69 | 85 | 95 | 125 | 187 |
| 6 | POLYMER.SCI | 669 | 59 | 72 | 44 | 43 | 47 | 58 | 73 | 63 | 107 | 103 |
| 7 | ENV.SCI | 606 | 17 | 21 | 33 | 29 | 51 | 43 | 47 | 94 | 100 | 171 |
| 8 | BIOTECH | 593 | 31 | 21 | 24 | 42 | 43 | 60 | 68 | 56 | 93 | 155 |
| 9 | PHYS:APPL | 560 | 15 | 12 | 39 | 27 | 44 | 49 | 45 | 88 | 93 | 148 |
| 10 | CHEM:PHYS | 527 | 22 | 41 | 32 | 34 | 49 | 35 | 55 | 59 | 55 | 145 |

ambiguity. First, homonymous names can occur whereby one distinct author keyword is associated to more than one individual. The datasets were constructed based on single subject categories in order to minimise such occurrences, that is, the likelihood for two or more people sharing the same nationality and subject category was assumed to be negligible.

Second, synonymous names may also occur in which an individual (typically one with a long sequence of initials) may identify him/herself differently from one publication to the next. Such occurrences can be detected by computing the character string similarity between pairs of author keywords. For our purposes, we used the Levenshtein edit distance and Jaro-Winkler distance due to their accuracy in detecting small variations in spelling, either deliberate or otherwise (Levenshtein 1966; Winkler 1990; Sariyar and Borg 2010). Both scores were computed on all possible[2] author keyword pairs contained within a given

---

[2] So as not to miss out any spelling variants.

subject category. A list of $N = 1,000$ authors, for example, will generate $N(N-1)$ $/2 \sim N^2 = 10^6$ author keyword pairs with scores varying in the range [0, 1] in which 0 and 1 denotes minimal and maximal similarity, respectively. We manually scrutinised author keyword pairs that generated a score above an acceptance threshold of 0.9 and above. If they were found to be interchangeable (refers to the same individual), we took this into account in our subsequent analyses.

Once authors were classified by nationality (local or foreign), the next step involves classifying the labels of local authors which depends on whether or not they have engaged in international collaboration. Recall that an international collaboration paper is identified as one where the authorship consists of at least one foreign author. Hence, a local author who has published at least one international collaboration paper is automatically classified as belonging to the international researcher sample. Likewise, an author with no track record of publishing with foreign coauthors (within the bounds of the study dataset), is assigned to the purely domestic researcher sample.

The same analysis was conducted on each of the 10 selected subject categories. By construction, since each dataset focuses on a single subject category, the article coverage of each dataset was effectively limited to source journals that comprised a given category. This means that contributions by purely domestic or international researchers outside of a given ISI subject category were necessarily excluded from analysis. This is a reasonable working assumption since we wish to discern impact within a narrow set of journals for which the focus and specialisation are similar and related to one another.

What we are asking then is this: "specific to the study period, does the configuration of Malaysian international researcher and purely domestic researchers produce similar impact within a given research field (as indicated by publication activity within journals listed under a specific ISI subject category, as well as citations associated to that activity)?" The following section describes our approach to answer precisely this question.

## Methods

We used a three-pronged methodology to compare the research performances between international and domestic research for specific subject categories in Malaysia. As seen in Fig. 1, our approach is composed of the following steps:

1. Conduct *distribution tests* to determine whether the citation distributions for DOM and INT were statistically identical. If such was the case, then no significant differential in impact could be assigned to either sample. In the non-identical case, the sample with the higher mean value was determined to have more impact (quality).
2. Compute *citation indices* to resolve any differences in the citation sequence.
3. Compute the *internationality index* $I(S)$ for each study dataset (corresponding to a single subject category) to determine the concentration of purely domestic researchers against their internationally linked counterparts.

The techniques used are discussed in more detail in subsections "Distribution tests", "Citation indices" and "Internationality index".
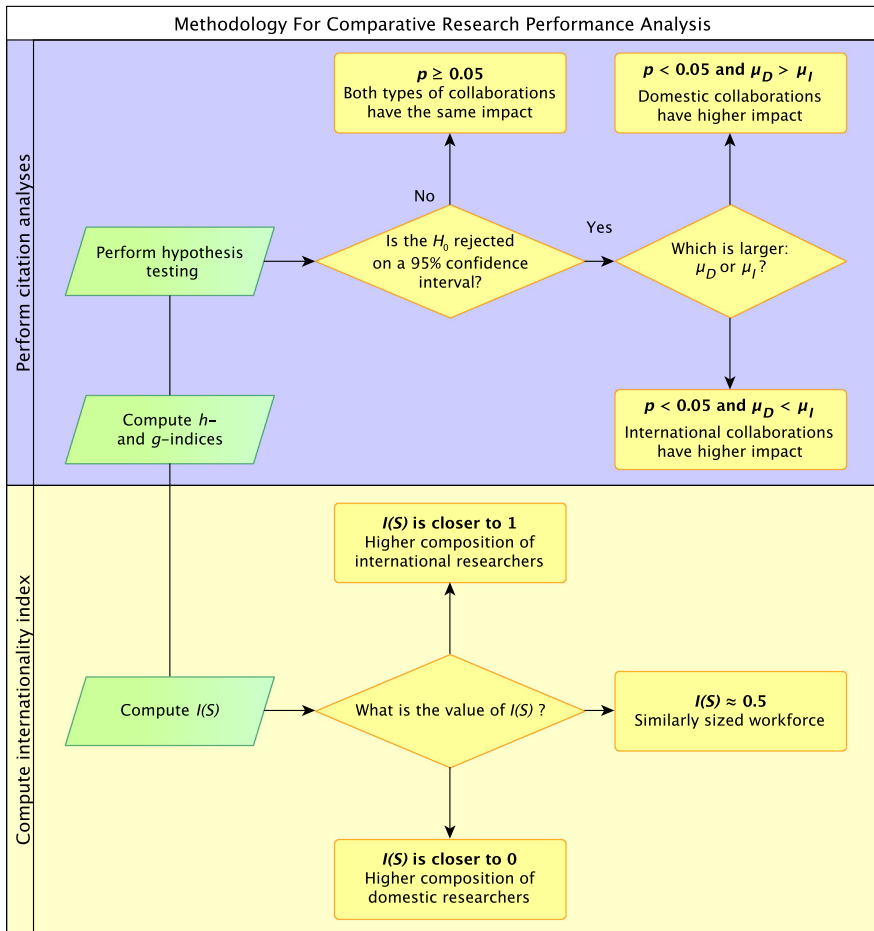
**Fig. 1** Outline of our methodology for comparative research performance analysis

## Distribution tests

The following statistical tests were used to determine if the citations distribution for DOM and INT are statistically identical: the Mann–Whitney–Wilcoxon (MWW) rank sum test and the bootstrap Kolmogorov–Smirnov (B-KS) test. Two statistical tests were used in this study to ensure mutual consistency. The usage of these tests were inspired by NK Avkiran's work on the empirical comparison of quality of collaborative research using the MWW test (Avkiran 1997).

Both tests have the advantage that they are non-parametric distribution tests; no assumptions are required as to the characteristic structure of the data or underlying population. However, the data must meet the following assumptions for both tests: (a) the data measurement scale should be of ordinal type i.e. the observations could be ranked or put in

order; and (b) all observations from both groups must must be independent of each other i.e. they come from distinct populations and do not affect each other.

We assume that the citation sequence for both DOM and INT are effectively i.i.d (independent and identically distributed) since the cumulative citation distribution for large sets of papers, $C(k) = \int_0^\infty P(k')\mathrm{d}k'$ (the probability at which a paper is cited at least $k$ times) tends to exhibit statistical regularities resembling some power-law $C(k) \sim k^{-v}\,(v > 0)$, stretched exponential $C(k) \sim \exp(-k^\beta)\,(\beta < 1)$, or log-normal form $C(k) = A\exp(-b\ln k - c(\ln k)^2)$ (Redner 2005).

For the citation sequence of a given sample (DOM or INT) to be i.i.d, the citation count for each paper in that sample must form a set of random variables $\{k_1, k_2, \ldots, k_n\}$ distributed according to some probability distribution $P(k)$ that governs the sample. That is, for a given sequence of papers, the number of citations received for each paper can be likened to a *times cited* (TC) dice roll[3] drawn from the same distribution $P(k)$. From the outset, we hypothesised that the citation distribution for DOM and INT are identical. One of the objectives of this paper is to test this hypothesis by means of the statistical tests described in subsections "Mann–Whitney–Wilcoxon test" and "Bootstrap Kolmogorov–Smirnov test".

### Mann–Whitney–Wilcoxon test

The MWW test is one of the more commonly used non-parametric tests which determines whether the data parameters (mean or median) for two independent samples are statistically identical.

This method is often called the Mann–Whitney U test, the MWW test or the Wilcoxon sum of ranks test as both Wilcoxon (1945), and Mann and Whitney (1947), independently worked out a non-parametric test using the same principles (Siegel and Castellan 1988).

As the citation data are highly skewed, the MWW test is preferable since it does not require the assumption of normality in the test data (Mann and Whitney 1947). The algorithm of the test procedure is displayed in Algorithm 1 (Nachar 2008).

---

**Algorithm 1:** Mann-Whitney-Wilcoxon test procedure

---

    **Consider**: Two independent samples, $F(x)$ and $G(y)$, denoted as $x_1, x_2, \ldots, x_{n_1}$, and $y_1, y_2, \ldots, y_{n_1}$ with sizes $n_1 < n_2$.

    **Null Hypothesis**: $H_0 : F(x) = G(y)$

1  Combine $x_i$ and $y_j$ samples.
2  Rank the combined samples in ascending order.
3  Divide the ranks into two sets according the group membership of the original observations: Let $r_1 i$ be the rank assigned to the first population, $x_i$ for $i = 1, 2, \ldots, n_1$, and $r_2 j$ be the rank assigned to the second population, $y_j$ for $j = 1, 2, \ldots, n_2$.
4  Calculate the test statistic: $U_x = n_1 n_2 + n_1(n_1 + 1)/2 - \sum_{i=1}^{n_i} r_1 i$; and $U_y = n_1 n_2 + n_2(n_2 + 1)/2 - \sum_{i=1}^{n_i} r_1 j$
5  For large sample sizes, $n_1, n_2 > 8$, a normal approximation is used: $z = (U - \mu_U)/\sigma_U$ where $\mu_U$ and $\sigma_U$ are the mean and standard deviation of $U$.
6  Reject $H_0$ if $p$-value of $\min(U_x, U_y) <$ significance level, $\alpha$.

---

---

[3] Each value in the sequence does not depend on previous values (i.e. each TC dice roll does not depend on previous rolls).

In this paper, there were 10 datasets prepared by subject category, each of which was split into DOM and INT samples. The median number of citations for both samples were compared (separately for each of the 10 datasets) using the MWW test to see if there was a significant difference based on a 95 % confidence interval (5 % degree of error). The significance test was done on the following null hypothesis.

$H_0$: There is no significant difference between the median number of citations for domestic and international collaboration articles, i.e., $\mu_D = \mu_I$.

Following Avkiran ([1997](#)), if the null hypothesis is rejected for a particular dataset (subject category) $S$ whereby the mean number of citations for INT papers is higher, then the implied conclusion is that international collaborations for Malaysian research in $S$ lead to better quality papers. Conversely, if the null hypothesis is rejected for a particular subject category $S$ whereby the mean number of citations for DOM papers is higher, then the implied conclusion is that domestic collaborations for Malaysian research in $S$ lead to better quality papers.

This was carried out using the `wilcox.test` function provided under the R package 'stats' (R Development Core Team [2008](#)).

*Bootstrap Kolmogorov–Smirnov test*

The Kolmogorov–Smirnov test (KS), on the other hand, is a widely used goodness-of-fit test that determines whether two independent samples were drawn from the same continuous distribution. This test was first introduced by Kolmogoroff ([1941](#)) and Smirnoff ([1939](#)).

A bootstrapped version of the univariate KS test (B-KS) was used in order to compute $p$ values in the presence of ties (Wang et al. [2003](#)). The advantage of this implementation is that it allows for comparisons between distributions that are not entirely continuous (a limitation of the regular KS test). The algorithm of the test procedure is displayed in Algorithm 2.

---

**Algorithm 2:** Bootstrapped Kolmogorov-Smirnov test procedure

---

    **Consider**: Two independent samples, $F(x)$ and $G(y)$, denoted as $x_1, x_2, \ldots, x_{n_1}$, and $y_1, y_2, \ldots, y_{n_1}$ with sizes $n_1 < n_2$. Also, $F_x(u)$ and $F_y(u)$ denote the values of the cumulative distribution functions at the point $u$ for the two samples, respectively.

    **Null Hypothesis**: $H_0 : F(x) = G(y)$

1  Simulated samples are randomly re-sampled from the original finite observed samples.
2  Calculate the test statistics (i.e. maximum difference in value) for every sample. Alternate hypothesis cases (i) $x < y$: $D^+ = max[F_x(u) - F_y(u)]$; (ii) $x > y$: $D^- = max[F_y(u) - F_x(u)]$
3  Select the optimal model whereby the maximum difference is closest to the mean difference of the whole simulated samples.
4  Reject $H_0$ if $p$-value $<$ significance level, $\alpha$.

---

Similar to the requirement for the MWW test, the null hypothesis is tested on a 95 % confidence interval. The null hypothesis corresponds to:

$H_0$: $x$ and $y$ have the same distribution ($F_x(u) = F_y(u)$ for all $u$).

This was carried out using the `ks.boot` function provided under the `R` package 'Matching' (Sekhon 2011). The number of bootstraps was set to 1,000.

Citation indices

The following quantitative bibliometric indicators were used to evaluate the research performance of a subject category: the Hirsch index and the Egghe index.

*Hirsch index*

The *h*-index is a single-number indicator that measures the quantity and impact of a given researcher (Hirsch 2005). It is a robust index that is insensitive to extreme values, such as lowly cited papers and highly cited papers. By Hirsch's definition, "a scientist has index *h* if *h* of his or her $N_p$ papers have at least *h* citations each and the other $(N_p - h)$ papers have $\leq h$ citations each".

Alternatively, given a sequence of *n* non-negative numbers $x = (x_1, \ldots, x_n)$, where $x_i \geq x_j$ for $i \leq j$, the *h*-index is defined as:

$$H(x) = \max\{i = 1, \ldots, n : x_i \geq i\}$$

if $n \geq 1$ and $x_1 \geq 1$, or $H(x) = 0$ otherwise (Gagolewski 2011).

One key disadvantage of the *h*-index is that low-productivity researchers with high impact papers may be unfairly assessed through this method as a researcher could only achieve the maximum value of their total number of publication (Costas and Bordons 2005; Egghe 2006; Moed 2000, 2005).

This was carried out using the `index.h` function provided under the `R` package 'CITAN' (Gagolewski 2011).

*Egghe index*

To counter the aforementioned drawback of the *h*-index, Egghe (2006) proposed a modification that takes into account the global citation performance of a set of papers: the *g*-index. By Egghe's definition, "the *g*-index is the highest rank such that the top *g* papers have, together, at least $g^2$ citations. This also means that the top $(g + 1)$ papers have less than $(g + 1)^2$ citations".

Alternatively, given a sequence of *n* non-negative numbers $x = (x_1, \ldots, x_n)$, where $x_i \geq x_j$ for $i \leq j$, the *g*-index is defined as:

$$G(x) = \max\left\{i = 1, \ldots, n : \sum\nolimits_{j=1}^{i} x_j \geq i^2\right\}$$

if $n \geq 1$ and $x_1 \geq 1$, or $G(x) = 0$ otherwise (Gagolewski 2011).

This was carried out using the `index.g` function provided under the `R` package 'CITAN' (Gagolewski 2011).

Internationality index

We proposed an internationality index *I(S)* to quantitatively gauge the concentration of domestic versus international researchers for a given *country-specific subject category* dataset *S*. Recall that in subsection "Author classification and disambiguation", a *purely*

*domestic researcher* is defined as a Malaysian author with no prior history of collaborating with a non-Malaysian coauthor within the study dataset. In contrast, an *international researcher* is defined as Malaysian author who has published at least one paper with foreign coauthors.

We can then define $I(S)$ as the fraction of international authors on the co-authorship network extracted from that dataset $(0 \leq I(S) \leq 1)$. The algorithm for our index is as displayed below.

---

**Algorithm 3:** Internationality Index algorithm

> **Input**: Co-authorship network $G = (V, E)$ for subject category $S$
> **Output**: Internationality index score $I(S)$
> **Require**: Locality of each node $v$ on $G$, $\lambda(v)$; $\lambda(v) = 1$ if researcher corresponding to node $v$ is *international* (i.e. $v$ has prior history of publishing international collaborative research); $\lambda(v) = 0$ otherwise (indicating that $v$ is *purely domestic*)
> 1 $G \leftarrow$ Extract co-authorship network of dataset $S$
> 2 $n \leftarrow$ Number of nodes on $G$
> 3 $n_I \leftarrow$ Number of international researchers on $G$
> 4 $I(S) \leftarrow n_I/n$

---

This is similar to Glänzel and Schubert's *authorship domesticity* score defined as "the ratio of the number of papers without international co-authorship [...] to the total number of papers of the given country" (Glänzel and Schubert 2005). There are two main differences here: first, we counted authors instead of papers; second, we used a network perspective as this paves the way for the construction of scoring algorithms based on the link structure between coauthors.

The co-authorship network $G = (V, E)$ is extracted based on author abbreviations associated to each article for a given dataset. Each article contributes a $k$-clique (a fully connected subgraph consisting of $k$ nodes) onto $G$. The overlap of nodes within each extracted $k$-clique increases the size of connected components on $G$ until the dataset is exhausted. The component with the largest number of nodes, termed the *giant connected component*, contains the most documented collaboration ties from which new, previously non-existent referrals can be traced.[4]

## Analysis

### Publication growth and distribution

The DOM and INT publication distributions and characteristics for Malaysia can be seen in Table 1. As a whole, Malaysian researchers have the propensity to collaborate with either international or domestic partners as opposed to producing single-authored articles between the years 2000–2009.[5] On average, 51.3 % of co-authored articles are

---

[4] Such referrals are an example of clustering by *choice homophily* (Kossinets and Watts 2009), which in turn, is due to *focal closure*. According to the theory of tie formation based on the confluence of "social interaction foci" known as *Focus Theory*, foci—consisting of various groups, contexts, and activities—organise and facilitate opportunities for interpersonal interactions (Feld 1981; McPherson et al. 2001). For the case of collaboration in research, focal closure describes the tendency of researchers to join or form communities or groups signifying specialisations on a particular field, topic, or sub-topic.

[5] All 949 single-authored articles from the years 2000–2009 were removed from the source data.

domestically co-authored with a range of 46.3–58.1 % throughout the study period. Internationally co-authored works, on the other hand, ranges from 41.9 to 53.7 % with a smaller annual mean percentage of 48.7 %. The top five countries involved in Malaysian international collaborations are China, India, England, Japan, and the USA.

The number of DOM publications shows a gradual improvement from 2002 onwards followed by a marked increase in the last 3 years. Similarly, INT publication numbers is also growing but at a slower pace. Collectively, Malaysian scientific output undergo a near six-fold increase (384–2,289 articles) and a near fivefold increase (355–1,653 articles) between 2000 and 2009, for DOM and INT, respectively. The *status quo* in Malaysian research and science is largely due to the numerous efforts made by the government since the 1990s to transform the country into a high-income knowledge-based economy. The marked increase in publications in the latter years is in line with the sharp rise in national R&D expenditures for the advancement of Research and Innovation based on the National Higher Education Strategic Plan (NHESP),[6] with an estimate of RM0.36 billion in 2006 to an estimate of RM2.5 billion in 2009 (MOSTI 2012).

Despite having lesser publications, INT articles, on average, receives greater numbers of citations than DOM articles annually. However, there is a diminishing average of citation values for both DOM and INT over the years; older articles generally garner more citations than recent work as citations accumulate over time.

As R&D expenditures and funding increase, scientific research becomes increasingly viable and these tend to small collaborations interspersed with rising amounts of large collaborations spanning several local institutions or countries. The number of DOM and INT authors and the corresponding average author values in Table 1 reflects this behaviour. The rise in the number of researchers and postgraduate students in public universities across the country (and resulting growth of domestic and international collaborations) after the introduction of NHESP is also mirrored by the high growth rate in the number of DOM and INT authors from 2007 and onwards.

As shown in Table 2, the top ten most productive subject categories each contributes around 3.1–13.6 % of Malaysia's total number of articles from 2000 to 2009. Crystallography (CRYSTAL) records the largest number of articles in Malaysia at 13.6 %, followed by Engineering, Electrical and Electronic (ENGN:ELEC) at 6.0 %, Material Science, Multidisciplinary (MAT.SCI) at 5.7 %, Food Science and Technology (FOOD.SCI) at 4.65 %, Engineering, Chemical (ENGN:CHEM) at 4.63 %, Polymer Science (POLYMER.SCI) at 4.3 %, Environmental Sciences (ENV.SCI) at 3.9 %, Biotechnology and Applied Microbiology (BIOTECH) at 3.6 %, Physics, Applied (PHYS:APPL) at 3.5 %, and finally, Chemistry, Physical (CHEM:PHYS) at 3.1 %.

In the following subsections, the differences in citation practices and researcher compositions across subject categories and how they impact international and domestic research are further examined.

Distribution test results

Non-parametric distribution tests—specifically, the MWW and B-KS tests—were used to determine the potentiality of impact for Malaysian research, whether lying equally on both

---

[6] The NHESP was launched in 2007 to establish Malaysia as an international and regional hub of excellence for higher education. Special emphasis is given to intensify research activities; aid the production of human capital and researchers; internationalise HEIs and research; and boost the development of S&T and R&D initiatives (MOHE 2007a, b).

domestic and international collaborative papers or, more on one type than the other. The test results indicate that the top ten datasets are split into two classes based on a significance level of $p < 0.05$: the first corresponding to the case $\mu(C_D) = \mu(C_I)$, and the second, the case where $\mu(C_D) < \mu(C_I)$. The results are displayed in Table 3.

For the *identical case* as shown in Table 3a, the strongest evidence (by decreasing $p$ value) appears in the FOOD.SCI dataset, followed by BIOTECH, ENGN:CHEM, and lastly, the POLYMER.SCI datasets. The last two datasets are interesting since their $\mu_D$ and $\mu_I$ values show the most difference, yet both the MWW and B-KS tests suggest that the domestic and international citation distributions are statistically identical. A common feature to be noted for these four subject categories is that the domestic samples are at least 1.8 times larger than the international ones (ENGN:CHEM), and at most four times larger (FOOD.SCI).

For the *non-identical case* as shown in Table 3b, there are six datasets where $\mu_D < \mu_I$. The CRYSTAL dataset is the only dataset with $N_I > N_D$. The remaining five datasets (ENGN:ELEC, MAT.SCI, ENV.SCI, PHYS:APPL, and CHEM:PHYS) are characterised by $N_D > N_I$, which suggests that international collaborative research sufficiently compensates lower productivity for impact. This may be due to halo effects from journals in which international collaborative papers are published in (if journal selection is indeed distinct from that in domestic papers), coverage differences in topic or subject matter (assuming these are associated with a marked citation rate differential), disparate composition of high impact researchers, more so when *coauthor citations*[7] received from a foreign researcher increases the visibility of the cited paper(s) (Merton 1968; Valderas 2007; Wuchty et al. 2007).

Internationality index scores

The internationality index score for a given subject category $S$ is computed as the fraction of the number of international researchers $n_I$, against the total number of researchers on the co-authorship network[8] of $S$, $n$.

Based on Table 4, the internationality index score, $I(S)$, for the *identical distribution case* lies within the range [0.36, 0.47]. In contrast, panel (b) shows that the *non-identical distribution case* covers the much wider range of [0.11, 0.84]. Of note, the CRYSTAL dataset yields $I(S) = 0.84$, which is possibly related to the larger volume of international compared to domestic collaborative papers ($N_D < N_I$). Here, the GCC of the CRYSTAL dataset accounts for 94 % of all researchers (as well as 96 % of their links) on the collaboration network. Furthermore, the GCC is found to be 129 times larger than the size of the second largest connected component (GCC2). Specific to the CRYSTAL dataset, a higher composition of international researchers is associated to higher productivity and larger mean impact.

The lowest score belongs to the ENGN:ELEC dataset, in which $I(S) = 0.11$ (i.e. 11 % of researchers on the co-authorship network $G$ have previously collaborated on an international level). Hence, the composition of researchers in ENGN:ELEC is primarily domestic, which coincides with higher domestic productivity $N_D = 607 > N_I = 313$, but

---

[7] Following Martin et al. (2013), *coauthor citations* refer to citations to a paper $X$ by researchers who have previously coauthored with any of the coauthors of $X$, yet are not themselves a coauthor to $X$. In contrast, *self-citations* refer to citations to $X$ by any of the coauthors to $X$, while *distant citations* refer to citations to $X$ by researchers who have had no prior or current history of collaborating with any of the coauthors of $X$.

[8] We assume that the co-authorship network is a sufficient proxy for the collaboration network.

**Table 3** Mann–Whitney–Wilcoxon (MWW) and Bootstrap Kolmogorov–Smirnov (B-KS) distribution test $p$ values and other citation characteristics for Malaysian top ten high-productivity subject categories. The null hypothesis $H_0$ is rejected for either test at $p < 0.05$

| No. | Subject category $S$ | MWW | B-KS | $N_D$ | $N_I$ | $C_D^{max}$ | $C_I^{max}$ | $\mu_D$ | $\mu_I$ | $M_D$ | $M_I$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (a) $H_0$: Identical distributions | | | | | | | | | | | |
| 4 | FOOD.SCI | 0.646 | 0.943 | 624 | 156 | 186 | 75 | 9.23 | 8.03 | 5 | 4 |
| 5 | ENGN:CHEM | 0.121 | 0.203 | 455 | 250 | 215 | 126 | 13.25 | 10.52 | 7 | 6 |
| 6 | POLYMER.SCI | 0.052 | 0.195 | 479 | 170 | 215 | 137 | 10.81 | 14.16 | 6 | 8 |
| 8 | BIOTECH | 0.387 | 0.689 | 415 | 169 | 145 | 60 | 11.06 | 9.69 | 6 | 6 |
| (b) $H_1$: Non-identical distributions | | | | | | | | | | | |
| 1 | CRYSTAL | 0.001 | 0.004 | 566 | 1713 | 85 | 69 | 2.96 | 3.43 | 1 | 2 |
| 2 | ENGN:ELEC | 0.000 | 0.000 | 607 | 313 | 166 | 197 | 5.43 | 7.53 | 2 | 3 |
| 3 | MAT.SCI | 0.000 | 0.000 | 499 | 261 | 114 | 188 | 7.68 | 10.50 | 4 | 7 |
| 7 | ENV.SCI | 0.000 | 0.002 | 348 | 235 | 149 | 345 | 11.87 | 14.65 | 6 | 8 |
| 9 | PHYS:APPL | 0.000 | 0.000 | 367 | 176 | 124 | 188 | 6.43 | 9.36 | 3 | 6 |
| 10 | CHEM:PHYS | 0.004 | 0.041 | 314 | 199 | 68 | 188 | 9.30 | 13.46 | 6 | 8 |

$N$, $C$, $\mu$, and $M$ denote total number of articles, citation count, mean citation count and median citation count for domestic ($D$) and international ($I$) samples, respectively (over the period 2000–2009)

**Table 4** The internationality index score and other network properties for Malaysian top ten high-productivity subject categories

| No. | Subject category, $S$ | $n$ | $m$ | $\chi$ | $n_D$ | $n_I$ | $p_{n_1}$ | $p_{m_1}$ | $r_{1,2}$ | $I(S)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CRYSTAL | 1,929 | 7,333 | 24 | 300 | 1629 | 93.7 | 96.0 | 129.1 | 0.84 |
| 2 | ENGN:ELEC | 7,179 | 1,8863 | 348 | 6356 | 823 | 81.7 | 88.9 | 136.4 | 0.11 |
| 3 | MAT.SCI | 2,363 | 6,070 | 159 | 1308 | 1055 | 63.8 | 70.0 | 33.5 | 0.45 |
| 4 | FOOD.SCI | 1,602 | 5,184 | 91 | 1031 | 571 | 72.2 | 82.1 | 52.6 | 0.36 |
| 5 | ENGN:CHEM | 1,373 | 3,357 | 92 | 730 | 643 | 69.1 | 77.7 | 55.8 | 0.47 |
| 6 | POLYMER.SCI | 982 | 2,500 | 52 | 565 | 417 | 73.3 | 77.2 | 19.5 | 0.42 |
| 7 | ENV.SCI | 1,955 | 8,154 | 201 | 775 | 1180 | 27.5 | 18.1 | 3.2 | 0.60 |
| 8 | BIOTECH | 1,664 | 5,595 | 109 | 930 | 734 | 63.5 | 71.9 | 34.0 | 0.44 |
| 9 | PHYS:APPL | 2,369 | 6,114 | 173 | 1703 | 666 | 64.9 | 73.4 | 43.9 | 0.28 |
| 10 | CHEM:PHYS | 1,313 | 3,887 | 90 | 582 | 731 | 59.9 | 71.6 | 31.5 | 0.56 |

$n$, $m$, $\chi$, $n_D$ and $n_I$ denote the number of nodes, number of links, number of components, number of purely domestic researchers, and number of international researchers, on the co-authorship graph $G$, respectively; $p_{n_1}$, $p_{m_1}$ and $r_{1,2}$ denote the percentage of nodes and links captured by the GCC, and the size of the GCC relative to GCC2, respectively; $I(S)$ denotes the internationality index score

lower mean impact $\mu_D = 5.43 < \mu_I = 7.53$ (see Table 3). This corresponds to an intriguing case of getting more from less. A similar case is apparent for PHYS:APPL as well, in which $I(S) = 0.28$ and $N_D = 367 > N_I = 176$, yet $\mu_D = 6.43 < \mu_I = 9.36$.

### Hirsch and Egghe indices and summarisation of results

The $h$- and $g$-indices are single-number indicators that measure the quantity and impact of a given researcher, or for the purpose of this paper, a given subject

category. The use of these indicators adds another dimension to our readings and understanding of research performance in Malaysia. A summarisation of the results for our three-pronged methodology for comparative research performance analysis between international and domestic research and the $h$- and $g$-indices for all ten datasets are displayed in Table 5.

In Table 5a, it can be seen that for the identical distribution case, the internationality index scores are typically $I(S) < 0.5$ indicating that the research composition is predominantly domestic. Although it follows from the B-KS and MWW tests that the citations for domestic and international articles are drawn from the same distribution, the $h$- and $g$-index values differ. Originally, the $h$- and $g$-indices are used to quantify the productivity and impact of an author. The above table adopts the same indices to show that while the potentiality of impact is similar, the realised citation sequences can differ between the domestic and international case. To a considerable extent, this is due to the relative size of the domestic versus international samples, wherein, $N_D/N_I \in [1.82, 4]$ for these four subject categories (FOOD.SCI, ENGN:CHEM, POLYMER.SCI, and BIOTECH). It then stands to reason that perhaps an increase in international researchers (corresponding to a boost in $I(S) \to 1$) leads to an increase in international collaborative papers ($N_I$). Indeed, it will be interesting to see if such a move triggers a shift in impact such that $\mu_D < \mu_I$.

In Table 5b, it can be seen that for the non-identical distribution case, the internationality scores show much more variety compared to the previous case. Recall that these six subject categories are classified as having $\mu_D < \mu_I$ according to the MWW test ($p < 0.05$). Hence, statistically speaking, international collaborative papers have higher impact compared to their domestic counterparts. Also interesting is the small gap between domestic and international $h$- and $g$-indices (with the exception of the CRYSTAL dataset since $N_D < N_I$ or $N_D/N_I = 0.33$). This suggests that although sample sizes are smaller for ENGN:ELEC, MAT.SCI, ENV.SCI, PHYS:APPL, and CHEM:PHYS, the international citation rates ($\mu_D < \mu_I$) sufficiently compensate for the lower level of productivity ($N_D > N_I$), which results in a smaller gap in Hirsch index ($h_D - h_I \to 0$).

A look at the Egghe indices in Table 5b however shows some interesting characteristics. For one thing, MAT.SCI exhibits $g_D < g_I$ which seems peculiar given that $h_D > h_I$ for that dataset. Furthermore, aside from the CRYSTAL dataset, all other datasets exhibit $g_D > g_I$ as well as $h_D \geq h_I$. This apparent discrepancy is possibly due to the inclusion of citation rates in the $g$-index formulation by Egghe. Hence, for MAT.SCI, the citation rate and sequence for international papers outweighs the domestic counterpart ($\mu_D = 7.68 < \mu_I = 10.50$; see Table 3) leading to $g_D < g_I$. There is also a similar (albeit more pronounced) pattern in the $g$-index values for CHEM:PHYS, which can perhaps be attributed to differences in the mean ($\mu_D = 9.30 < \mu_I = 13.46$; see Table 3), as well.

## Discussion

As Malaysia vies to transform its nation into a more knowledge-intensive economy, there have been many improvements made when it comes to research in terms of funding, human capital development, infrastructure, and even in networking intensifying activities between domestic and international researchers to stimulate knowledge and technology transfer. Malaysia has also incorporated internationalisation into their strategic plans in

**Table 5** Summarisation of results in Tables 3 and 4, followed by the Hirsch ($h$) and Egghe ($g$) index values

| No. | Subject category, $S$ | $N_D$ | $N_I$ | $C_D^{max}$ | $C_I^{max}$ | $M_D$ | $M_I$ | $I(S)$ | $h_D$ | $h_I$ | $g_D$ | $g_I$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (a) $H_0$: Identical distributions | | | | | | | | | | | | |
| 4 | FOOD.SCI | 624 | 156 | 186 | 75 | 5 | 4 | 0.36 | 33 | 18 | 47 | 26 |
| 5 | ENGN:CHEM | 455 | 250 | 215 | 126 | 7 | 6 | 0.47 | 35 | 25 | 53 | 36 |
| 6 | POLYMER.SCI | 479 | 170 | 215 | 137 | 6 | 8 | 0.42 | 31 | 26 | 44 | 39 |
| 8 | BIOTECH | 415 | 169 | 145 | 60 | 6 | 6 | 0.44 | 33 | 21 | 47 | 28 |
| (b) Non-identical distributions | | | | | | | | | | | | |
| 1 | CRYSTAL | 566 | 1713 | 85 | 69 | 1 | 2 | 0.84 | 14 | 22 | 25 | 32 |
| 2 | ENGN:ELEC | 607 | 313 | 166 | 197 | 2 | 3 | 0.11 | 24 | 22 | 40 | 36 |
| 3 | MAT.SCI | 499 | 261 | 114 | 188 | 4 | 7 | 0.45 | 25 | 24 | 35 | 38 |
| 7 | ENV.SCI | 348 | 235 | 149 | 345 | 6 | 8 | 0.60 | 30 | 30 | 46 | 45 |
| 9 | PHYS:APPL | 367 | 176 | 124 | 188 | 3 | 6 | 0.28 | 21 | 18 | 32 | 30 |
| 10 | CHEM:PHYS | 314 | 199 | 68 | 188 | 6 | 8 | 0.56 | 25 | 25 | 33 | 40 |

order to attain recognition and competitive advantage on a national and global scale[9]. However, more effort should be put into facilitating effective and productive scientific research collaboration on a domestic level, and more so on an international level.

As recommended by the Internationalisation Policy, Malaysia should endeavour to identify niche areas to focus on in their R&D activities and also form strategic partnerships that would be beneficial to research productivity (MOHE 2011). With our methodology, we can help narrow down the range of subject categories (or research areas) that would benefit most from increased partnerships in international (or domestic) collaboration. Thus, certain key areas should be identified and prioritised.

Policy and decision making should be backed by clear, rational, and effective assessment guidelines. In this sense, the process of gauging a country's researcher composition, whether leaning more towards domestic or international researchers, is a fairly straightforward process that only requires information on the co-authorship of each paper in the target dataset $S$ and the locality of each author, $\lambda$ (see Algorithm 3). The resulting internationality score $I(S) = n_I/n_1$ is bounded within the range [0, 1], in which $I(S) \to 1$ indicates a higher international composition, while $I(S) \to 0$ indicates a higher domestic composition. If increased international composition is set as a national milestone to be achieved, $I(S)$ can be used as a key performance indicator (KPI) to determine whether a given trajectory is on or off target.

Consider instead the impact assessment of domestic and international research. Here, citation counts are used as a proxy for impact. Given a specific year range $[t_\alpha, t_\Omega]$ and subject category $S$, non-parametric distribution tests like those implemented in this paper can be used to determine whether impact is equal or more biased on one sample than the other (domestic or international collaborative papers). If one is more dominant than the other, the question then becomes *how much so*? In this, the mean citation count ($\mu$) or

---

median citation count ($\mu_{1/2}$ or $M$) can be seen as population estimates. The use of sample means however is sensitive to extreme values, which makes sample medians a better option since it circumvents this issue while accounting for the split (midpoint) in the data.

There are other viable and potentially more insightful options. In particular, Hirsch ($h$) and Egghe ($g$) indices (Hirsch 2005; Egghe 2006) can be used to estimate the level of impact, as well as the productivity, associated to domestic and international samples of each study dataset. More importantly, by separating Hirsch and Egghe index scores by their domestic and international components ($h_D$, $h_I$ and $g_D$, $g_I$), the size of the gap between domestic and international collaborations can be gauged. This should give policy and decision makers the means to assess the present state of research, which can then be used to set feasible milestones, as demonstrated in section "Hirsch and Egghe indices and summarisation of results".

## Conclusion

Our analysis shows that Malaysian research at large benefits from both types of scientific collaborations in varying degrees: though a higher percentage of ISI articles in Malaysia is credited to domestic publications, international publications as a whole is attributed to higher annual citation counts and average citation values throughout the study period of 2000–2009. This trend is repeated in the top ten subject categories.

This points to two things: firstly, domestic scientific partnerships are the backbone to top tier Malaysian research; secondly, though increasing the number of international collaborators is not a guarantee of increased citation impact, there is a greater likelihood that papers with foreign affiliations will accumulate a higher amount of citations than domestic-only co-authored papers. In principle, the larger the number of authors to a given paper, the more self-citations or coauthor citations it can generate,[10] especially if any one of these authors or their coauthors splinters off to follow up on the work independently from the rest of the original team (Valderas 2007). We expect this scenario to be more likely for international collaborations.

The results of the analysis presented in this paper suggests a need for strategies that will improve the visibility of domestic collaborations as well as the implementation of measures to intensify collaborations with foreign partners. Malaysia must take careful stock of the dynamics of knowledge production on both levels to ensure that the quality of its publications are on par with major competitors on the world stage. Furthermore, the analysis carries useful implications for decision-makers and stakeholders involved in research policy planning and management in Malaysia. The methodology used in this paper may also prove beneficial to the formulation of best practices for research collaboration in other developing nations.

---

[10] *Self-citation* occurs when a paper $B$ refers to paper $A$, given that at least one author in $B$ is also a member in $A$. In contrast, a *coauthor citation* occurs when a paper $Y$ refers to paper $X$, given that at least one author in $X$ has coauthored with another in $Y$, yet authorship in $X$ and $Y$ itself does not overlap (Martin et al. 2013).

# References

Aguillo, I. F., Bar-Ilan, J., Levene, M., & Ortega, J. L. (2010). Comparing university rankings. *Scientometrics*, *85*(1), 243–256.

Anuradha, K. T., & Urs, S. R. (2007). Bibliometric indicators of indian research collaboration patterns: A correspondence analysis. *Scientometrics*, *71*(2), 179–189.

Avkiran, N. (1997). Scientific collaboration in finance does not lead to better quality research. *Scientometrics*, *39*(2), 173–184.

Basu, A., & Kumar, B. V. (2000). International collaboration in indian scientific papers. *Scientometrics*, *48*(3), 381–402.

Beaver, D. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics*, *52*(3), 365–377.

Beaver, D., & Rosen, R. (1978). Studies in scientific collaboration. *Scientometrics*, *1*(1), 65–84.

Beaver, D., & Rosen, R. (1979a). Studies in scientific collaboration-part II. Scientific co-authorship, research productivity and visibility in the french scientific elite, 1799–1830. *Scientometrics*, *1*(2), 133–149.

Beaver, D., & Rosen, R. (1979b). Studies in scientific collaboration part III. Professionalization and the natural history of modern scientific co-authorship. *Scientometrics*, *1*(3), 231–245.

Buela-Casal, G., Gutiérrez-Martínez, O., Bermúdez-Sánchez, M. P., & Vadillo-Muñoz, O. (2007). Comparative study of international academic rankings of universities. *Scientometrics*, *71*(3), 349–365.

Campbell, D., Côté, G., Haustein, S., Lefebvre, C., & Roberge, G. (2014). *Bibliometric study in support of Norway's strategy for international research collaboration*. Canada: Science-Metrix.

Castells, M. (1996). *The information age: Economy, society and culture: The rise of the network society* (Vol. 1). Cambridge, MA: Blackwell Publishers Inc.

Cave, M. (1997). *The use of performance indicators in higher education: The challenge of the quality movement*. London: Jessica Kingsley Publishers.

Cheng, M. Y., Hen, K. W., Tan, H. P., Fok, K. F. (2013). Patterns of co-authorship and research collaboration in malaysia. In Aslib Proceedings, volume 65, pages 659–674. Emerald Group Publishing Limited.

Costas, R., & Bordons, M. (2005). Bibliometric indicators at the micro-level: Some results in the area of natural resources at the Spanish CSIC. *Research Evaluation*, *14*(2), 110–120.

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press.

Davarpanah, M. R. (2009). The international publication productivity of malaysia in social sciences. *Journal of Scholarly Publishing Journal of Scholarly Publishing*, *41*(1), 67–91.

Egghe, L. (2006). An improvement of the H-index: The G-index. *ISSI Newsletter*, *2*(1), 8–9.

Feld, S. (1981). The focused organization of social ties. *American Journal of Sociology*, *86*(5), 1015–1035.

Gagolewski, M. (2011). Bibliometric impact assessment with R and the CITAN package. Journal of Informetrics.

Garfield, E. (1963). Citation indexes in sociological and historical research. *American Documentation*, *14*(4), 289–291.

Garfield, E., et al. (1964). Science citation index—a new dimension in indexing. *Science*, *144*(3619), 649–654.

Georghiou, L. (1998). Global cooperation in research. *Research Policy*, *27*(6), 611–626.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, *51*(1), 69–115.

Glänzel, W., & Schubert, A. (2005). Domesticity and internationality in co-authorship, references and citations. *Scientometrics*, *65*(3), 323–342.

González-Albo, b, & Bordons, M. (2011). Articles vs. proceedings papers: Do they differ in research relevance and impact? A case study in the library and information science field. *Journal of Informetrics*, *5*(3), 369–381.

Griffith, B. C., Jahn, M. J., & Miller, A. J. (1971). Informal contacts in science: A probabilistic model for communication processes. *Science*, *173*(3992), 164–166.

Griffith, B. C., Small, H. G., Stonehill, J. A., & Dey, S. (1974). The structure of scientific literatures II: Toward a macro-and microstructure for science. *Social Studies of Science*, *4*(4), 339–365.

He, T. (2009). International scientific collaboration of china with the G7 countries. *Scientometrics*, *80*(3), 571–582.

Heffner, A. G. (1981). Funded research, multiple authorship, and subauthorship collaboration in four disciplines. *Scientometrics*, *3*(1), 5–12.

Heimeriks, G., & Vasileiadou, E. (2008). Changes or transition? Analysing the use of ICTs in the sciences. *Social Science Information*, *47*(1), 5–29.

Heinze, T., & Kuhlmann, S. (2008). Across institutional boundaries?: Research collaboration in German public sector nanoscience. *Research Policy*, *37*(5), 888–899.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569.

Katz, J., & Martin, B. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18.

Katz, J. S., & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, *40*(3), 541–554.

Kaur, S., Ibrahim, R., Selamat, A. (2013). Constraints on achieving key performance indicators for scholarly publications among academic staff: Case of a malaysian public university. In Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on, pages 73–78. IEEE.

Kolmogoroff, A., et al. (1941). Confidence limits for an unknown distribution function. *The Annals of Mathematical Statistics*, *12*(4), 461–463.

Kossinets, G., & Watts, D. J. (2009). Origins of homophily in an evolving social network. *American Journal of Sociology*, *115*(2), 405–450.

Kumar, S., & Jan, J. M. (2013). Mapping research collaborations in the business and management field in malaysia, 1980–2010. *Scientometrics*, *97*(3), 491–517.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Cybernetics and Control Theory*, *10*(8), 707–710.

Low, W. Y., Ng, K. H., Kabir, M. A., Koh, A. P., & Sinnasamy, J. (2014). Trend and impact of international collaboration in clinical medicine papers published in malaysia. *Scientometrics*, *98*(2), 1521–1533.

Malairaja, C., & Zawdie, G. (2008). Science parks and university–industry collaboration in malaysia. *Technology Analysis and Strategic Management*, *20*(6), 727–739.

Mann, H. B., Whitney, D. R., et al. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60.

Martin, T., Ball, B., Karrer, B., Newman, M. (2013). Coauthorship and citation in scientific publishing. arXiv preprint corr/abs-1304-0473.

McPherson, M., Smith-Lovin, L., & Cook, J. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444.

Melin, G. (2000). Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy*, *29*(1), 31–40.

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63.

Miquel, J.-F., & Okubo, Y. (1994). Structure of international collaboration in science-Part II: Comparisons of profiles in countries using a link indicator. *Scientometrics*, *29*(2), 271–297.

Moed, H. (2005). *Citation analysis in research evaluation, volume 9*. Berlin: Springer.

Moed, H. F. (2000). Bibliometric indicators reflect publication and management strategies. *Scientometrics*, *47*(2), 323–346.

MOHE. (2007a). *National higher education action plan: 2007–2010*. Ministry of Higher Education: Putrajaya.

MOHE. (2007b). *Pelan strategik pengajian tinggi negara: Melangkaui tahun 2020 (The national higher education strategic plan: Laying the foundation beyond 2020)*. Ministry of Higher Education: Putrajaya.

MOHE. (2011). *Internationalisation policy for higher education Malaysia 2011*. Ministry of Higher Education: Putrajaya.

MOSTI. (2012). National survey of research and development (R&D) Malaysia. Ministry of Science, Technology and Innovation, Malaysia: Putrajaya.

Nachar, N. (2008). The Mann–Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, *4*(1), 13–20.

National Science Board. (2012). *Science and engineering indicators 2012*. National Science Foundation: Arlington, VA.

Niu, F., & Qiu, J. (2014). Network structure, distribution and the growth of chinese international research collaboration. *Scientometrics*, *98*(2), 1221–1233.

Price, D. J. S. (1963). *Big science, little science*. New York: Columbia University.

Price, D. J. S., & Beaver, D. (1966). Collaboration in an invisible college. *American Psychologist*, *21*(11), 1011.

R Development Core Team. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna. ISBN 3-900051-07-0.

Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*, *58*, 49–54.

Sariyar, M., & Borg, A. (2010). The record linkage package: Detecting errors in data. *The R Journal*, *2*(2), 61–67.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, *42*(7), 1–52.

Siegel, S., Castellan, N. (1988). Nonparametric statistics for the behavioral sciences.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269.

Small, H., Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. Science Studies, pp. 17–40.

Smirnoff, N. (1939). Sur les écarts de la courbe de distribution empirique. *Matematicheskii Sbornik*, *48*(1), 3–26.

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*(1), 643–681.

Thorsteinsdóttir, O. H. (2000). External research collaboration in two small science systems. *Scientometrics*, *49*(1), 145–160.

Valderas, J. (2007). Why do team-authored papers get cited more? *Science (New York, NY)*, *317*(5844), 1496.

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, *34*(10), 1608–1618.

Wang, J., Tsang, W. W., & Marsaglia, G. (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, *8*(18).

Wang, X., Xu, S., Wang, Z., Peng, L., & Wang, C. (2013). International scientific collaboration of China: Collaborating countries, institutions and individuals. *Scientometrics*, *95*(3), 885–894.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. In *Proceedings of the section on survey research methods*. American Statistical Association, ERIC, pp. 354–369.

Wuchty, S., Jones, B., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039.

Zhou, P., & Glänzel, W. (2010). In-depth analysis on china's international cooperation in science. *Scientometrics*, *82*(3), 597–612.