Applying the ecological Shannon's diversity index to measure research collaboration based on coauthorship: A pilot study

Ari Voutilainen*, Mari Kangasniemi

Department of Nursing Science, University of Eastern Finland, Kuopio, Finland

ABSTRACT

The purpose of this study was to test the usefulness of a slightly modified Shannon's diversity index (H) as a numerical measure of intragroup research collaboration diversity based on coauthorship. Altogether, 527 peer-reviewed scientific papers by two university departments were used as the study material. Nonrandom rationalized sampling was executed to enable the confirmation of the authors' affiliations. The smallest unit of collaboration, i.e., a pair of authors, was created by matching every author with each of the coauthors from the same department he or she collaborated with. H was calculated at the department level and compared with the previously published, coauthorship based measures of research collaboration: The collaborative index (CI), degree of collaboration (DC) and collaboration diversity index (CDI). Obviously, H expressed a different aspect of research collaboration than the existing indexes. Compared to CI, DC, and CDI, H revealed novel aspects of collaboration when the abundance of collaboration increased and the distribution of collaborative relations between coauthors moved closer to the uniform distribution at the same time. H can provide additional information about collaborative relationships between researchers based on coauthorship, and it should be considered as a partial indicator of research collaboration.

Keywords: Coauthorship, diversity, research collaboration

INTRODUCTION

Research collaboration is a complex phenomenon with several aims. These include increasing publishing productivity and the number of citations, publishing in journals with high impact factors, minimizing risks and maximizing opportunities for single researchers and expanding the base of knowledge and producing economic value.[1-5] The quality, form, process, costs, and motivations of collaboration have been studied and reported over the last few decades. The management of heterogeneous research groups,[4] different research cultures,[6] social structures,[7] such as a sense of community,[8] and the characteristics and profiles of the collaborators,[5,9] have been recognized as important factors for collaboration [Figure 1]. Exploring how to build collaborative teams has been found to be crucial for the success of the collaboration.[4]

Research collaboration among individuals, groups, departments, institutions, sectors and countries[5,10] has become the norm in every field of scientific research.[1,4,5] It has been seen as a goal in itself, but also as a tool to

***Address for correspondence:**
E-mail: ari.voutilainen@uef.fi

Access this article online

Quick Response Code:

Website:
www.jscires.org

DOI:
10.4103/2320-0057.174866

How to cite this article: Voutilainen A, Kangasniemi M. Applying the ecological Shannon's diversity index to measure research collaboration based on coauthorship: A pilot study. J Sci Res 2015;4:172-7.

increase research productivity and efficiency as well as its effectiveness. Scientific productivity [Figure 1], such as impact factors and citations counts,[11] has been presented as a way to measure the results of research collaboration.[3] In particular, the positive effect of collaboration on scientific productivity has been highlighted by Lee and Bozeman,[1] Liao and Yen,[3] Brew *et al.*,[4] and Katz and Martin.[10]

Coauthorship has proved a useful way of symbolizing research collaboration,[3,5,10] although many counterarguments have also been proposed.[5] Different indexes based on different computations of coauthorships have been developed to illustrate research collaboration.[12,13] These include, for example, the collaboration diversity index (CDI),[14] the collaborative index (CI),[15] the degree of collaboration (DC)[16] and the collaborative coefficient (CC).[17] Briefly, CDI is the number of unique coauthors divided by the total number of collaborative relationships at the author level, CI describes the average number of authors per paper for a given set of papers, DC is the fraction of multiple-authored papers and CC combines CI and DC [Figure 1].

Despite the multifaceted nature of research collaboration, no attention has been paid to the intragroup diversity of collaborative relationships between coauthors [Figure 1]. Diversity is one of the fundamental measures in ecology that denotes the even distribution and abundance of species.[18-20] Diversity is high when there are plenty of different species present in the community, and the proportions of species do not differ from each other. Thus, diversity, in general, is an informative variable and to date, there is no measure that simultaneously reflects both the evenness and richness of collaborative relationships among coauthors.

The purpose of this study was to measure intragroup collaboration diversity between coauthors by testing a simple method, the Shannon's diversity index (H),[18,19] and to compare it to the previously established measures of research collaboration: CDI, CI, and DC. To achieve our purpose, H was slightly modified and associated with the number of citations per publication. We suggest that H measures different parameters of research collaboration to CDI, CI, DC, and CC. It also is assumed that diverse collaboration becomes desirable if it is pointed out that diversity increases the number of citations per publication, and vice versa if diverse collaboration leads to fewer citations.

SUBJECTS AND METHODS

Empirical Application

The present example was based on the coauthorship of scientific papers and number of citations per paper. The first was considered to reflect part of the research collaboration and the latter to reflect scientific productivity. The purpose was not to evaluate the resulting H values *per se*, whether they were high or low, but to demonstrate that there could be an alternative way to express research collaboration, especially since the operationalization of the concept has proved to be a difficult task. H was compared with the existing measures of collaboration, namely CDI, CI, and DC, to test our suggestion that H creates novel information about research collaboration

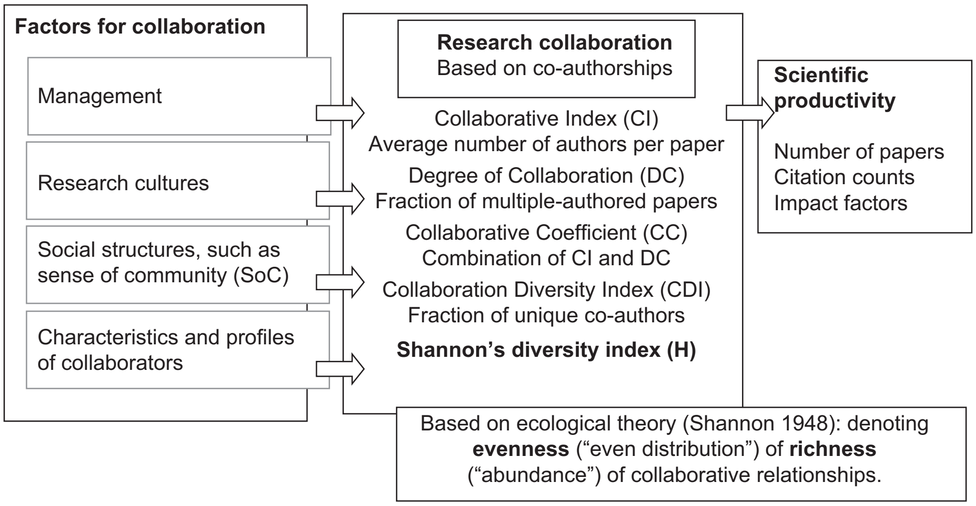
Factors for collaboration
Management
Research cultures
Social structures, such as sense of community (SoC)
Characteristics and profiles of collaborators
Research collaboration
Based on co-authorships
Collaborative Index (CI)
Average number of authors per paper
Degree of Collaboration (DC)
Fraction of multiple-authored papers
Collaborative Coefficient (CC)
Combination of CI and DC
Collaboration Diversity Index (CDI)
Fraction of unique co-authors
Shannon's diversity index (H)
Scientific productivity
Number of papers
Citation counts
Impact factors
Based on ecological theory (Shannon 1948): denoting **evenness** ("even distribution") of **richness** ("abundance") of collaborative relationships.

Figure 1: Shannon's diversity index (H) as a part of research collaboration based on coauthorship

based on coauthorship at a group-level. CC was excluded from the comparison, as it is not an independent measure but a combination of CI and DC.

Data

The data were comprised scientific papers produced by two university departments, one focusing on health sciences and the other on natural sciences. In the analyses, the departments were kept separate to enable independent verification of the calculations. All the papers that were included were peer-reviewed and published in scientific journals between 2007 and 2012. Altogether, 527 publications were gathered between April 22 and July 12, 2013, from open access databases maintained by the universities. Nonrandom rationalized sampling was executed in this pilot study to allow the confirmation of the authors' affiliations. During the first phase, the names and numbers of authors of the publications were extracted and tabulated. The names were then replaced with codes, for ethical reasons, and the authors were randomly numbered while still retaining their connections to the original publications. Data matrices were created for each department. The authors of peer-reviewed scientific publications were classified as those representing the departments in question and those representing other departments and organizations on the basis of their main affiliations. The latter will henceforth be called companion authors. To be able to associate the publication data with scientific productivity, the number of citations per publication was obtained from Google Scholar on November 28, 2014 (Health Sciences Department) and on December 8, 2014 (Natural Sciences Department). Google Scholar was chosen as the source because it appeared to be more up-to-date than other citation databases.

Shannon's Diversity Index

H was based on the number of coauthorships between the authors representing the department in question, namely health or natural sciences. Each combination of coauthors for each paper was treated as a pair. This in turn meant that the theoretical maximum number of pairs per department $n_{\max}$ was N-1, when the total number of authors per department (N) was 2, N-1 + N-2 when N was 3, N-1 + N-2 + N-3 when N was 4 and so on. A pair of authors was chosen as the study unit instead of a single author, as a single author cannot be studied for collaboration. In other words, a pair of authors was, implicitly, the smallest unit of collaboration. As the aim was to study collaboration diversity within a group and the groups of interest were *a priori* assigned, the companion authors were omitted from the calculations. In other words, H indicated the level of "intragroup", not "intergroup" collaboration. The equation was:

$$\mathrm{H} = \frac{-\sum_{i=1}^{n} p_i \log_{10} p_i}{\log_{10} \frac{\mathrm{N}(\mathrm{N}-1)}{2}} \qquad (1)$$

Where p_i was the proportion of collaboration carried out by a particular pair of authors of all pair-wise collaborations, n was the number of actualized pairs of authors, not $n_{\max}$, and N was the total number of authors from the department in question. H favors evenness; that is, uniform distribution of collaboration across the pairs of authors. This means that two collaborations between the authors *i* and *i'* increases H less than one collaboration between *i* and *i'* summed to one collaboration between *i* and *i"* (2*ii'* <*ii'* + *ii"*). It is crucial to note that H is not a measure of asymmetry in the same way as the adjusted Fisher-Pearson standardized moment coefficient, also known as skewness.[20,21]

The resulting H values can be presented as such or they can be normalized. The raw H value has no upper limit and describing it with adjectives such as low and high is ambiguous. Thus, normalization is suggested if the purpose is to compare the H values across time periods and/or groups. A simple way to normalize the index values to a range of between 0 and 1 within the group of interest is as follows:

$$\frac{\mathrm{H}_i - \mathrm{H}_{min}}{\mathrm{H}_{\max} - \mathrm{H}_{\min}} \qquad (2)$$

Where *i* refers to time periods, such as years from 2007 to 2012 in the present case, and corresponds to the number of index values to be reported. If the goal is to compare the indexes across different groups, the best way to help the interpretation is to normalize the values over the groups. This method of interpretation is flexible and can be modified, depending on what the researcher aims to achieve.

Control information was produced by (i) calculating correlations across H, CDI, CI, DC, and the scientific productivity, namely the number of citations per publication, and (ii) plotting the productivity to primary data, namely the number of authors per publication. The plots were drawn separately for the authors from the departments of interest and for the companion authors.

RESULTS

The annual nonnormalized H ranged between 0.703 and 0.888 in the Health Sciences Department and between 0.689 and 0.788 in the Natural Sciences Department [Figure 2]. The lack of strong correlations across CI, DC, CDI, and H demonstrated that the indexes were independent so that they were measuring different aspects of research collaboration based on coauthorships [Figure 2 and Table 1]. The direction of the relationship between the number of citations per publication and measure of research collaboration differed across the indexes [Table 1].

The number of citations per publication appeared to relate to the number of authors per publication, so that more authors from the department in question decreased the number of citations, whereas the companion authors increased the number of citations, especially in the Natural Sciences Department [Figure 3]. The number of departmental authors per publication correlated negatively with the number of companion authors, both in the Health and Natural Sciences Departments (Spearman's ρ = −0.620 and −0.371, respectively). However, the effects of the department and companion authors on the number of citations were not clearly transmitted via the association between the numbers of department and companion authors, as the number of citations was only weakly correlated with the ratio of companion to department authors (Health: ρ =0.054; Natural: 0.209) [Figure 3].

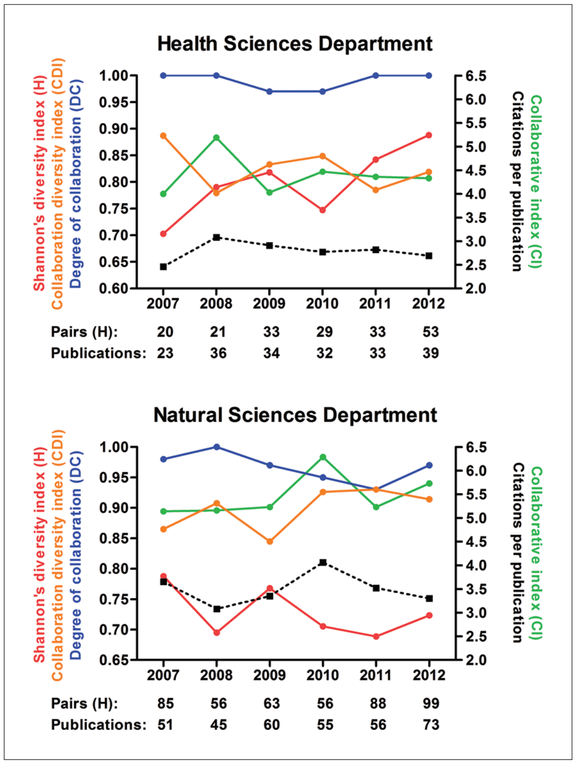

Figure 2: Annual Shannon's diversity index (H, red colour), collaboration diversity index collaboration diversity index (CDI, orange), degree of collaboration (DC, blue), and collaborative index (CI, green), and the mean number of citations per publication (black) in the two university departments

DISCUSSION AND CONCLUSIONS

It was obvious that H expressed a different aspect of research collaboration based on coauthorship than the existing indexes, CI and DC. On the other, H related negatively to CDI in the case of the Natural Sciences Department, so that when CDI was high, H was low and vice versa [Figure 2]. This was most probably due to the presentation of the results. Over the course of a year, researchers from the natural science department had published with many unique companion coauthors (high CDI) and collaborated less with potential coauthors from their department (low H). If the data had been pooled over the years, the result would have changed as the number of unique coauthors would have decreased and the number of collaborations between the department authors would have increased. In other words, researchers from the natural science department collaborated with fewer unique coauthors over the study period, 2007–2012, but they only collaborated with each coauthor once a year. In the case of the Health Sciences Department, H did not follow CDI because researchers from the health

Table 1: Nonparametric correlations (Spearman's ρ) between the collaborative index, degree of collaboration, collaboration diversity index, Shannon's diversity index, and the number of citations per publication per year

Department	CI	DC	CDI	H	Citations
Health sciences					
CI	1	0.000	−0.657	0.086	0.600
DC	0.000	1	−0.414	0.207	−0.207
CDI	−0.657	−0.414	1	−0.600	−0.714
H	0.086	0.207	−0.600	1	0.200
Citations	0.600	−0.207	−0.714	0.200	1
Natural sciences					
CI	1	−0.662	0.522	−0.261	0.232
DC	−0.662	1	−0.696	0.406	−0.522
CDI	0.522	−0.696	1	−0.771	0.257
H	−0.261	0.406	−0.771	1	0.200
Citations	0.232	−0.522	0.257	0.200	1

All correlations were statistically nonsignificant (*P*>0.05 in each case). CI=Collaborative index, DC=Degree of collaboration, CDI=Collaboration diversity index, H=Shannon's diversity index

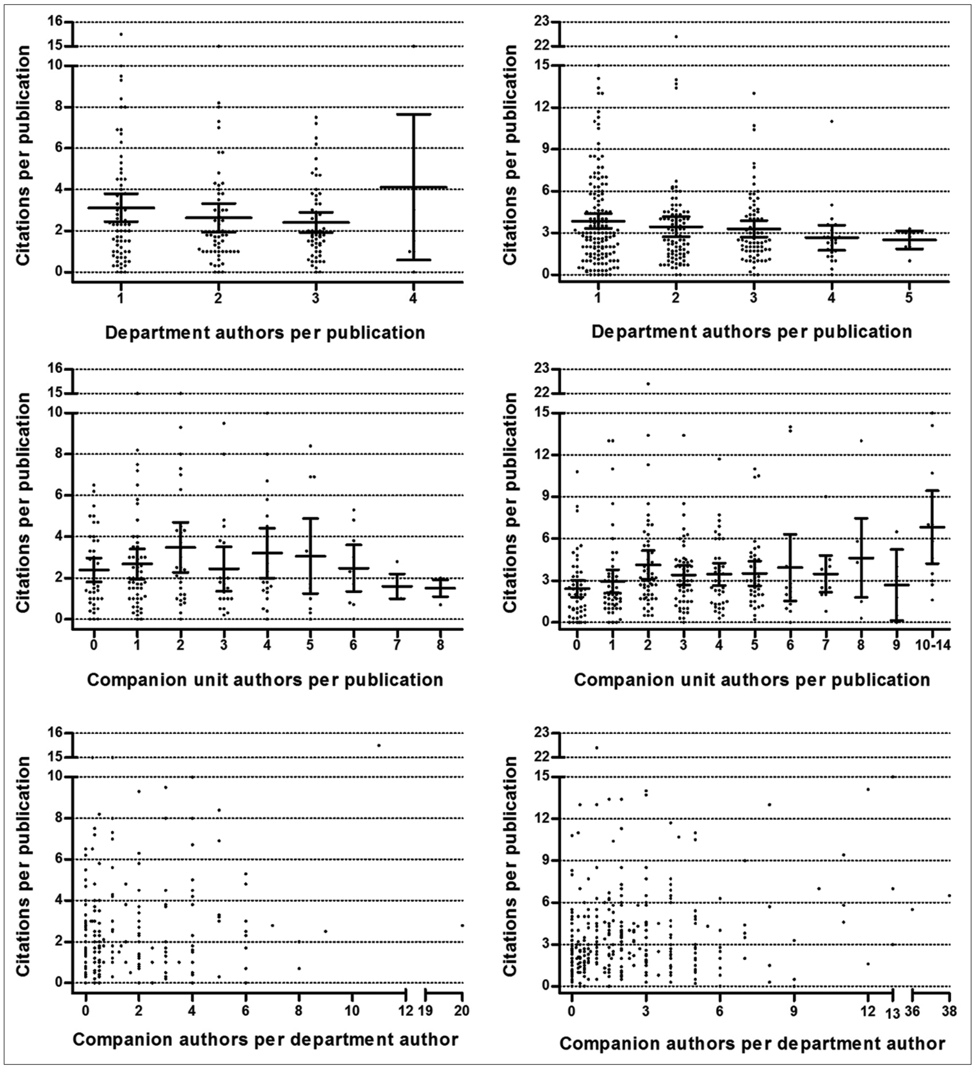

Figure 3: Relationships between the number of citations and authors per publication in the Health (left column) and Natural Sciences Department (right column). The symbol indicates mean and 95% confidence interval. Each dot represents one publication

sciences unit had collaborated more with coauthors from their own department and/or many times with the same coauthors a year. This finding emphasizes different application possibilities of the two collaboration diversity measures: CDI generated by Li *et al.*[14] and the slightly modified H launched in this paper. CDI is suitable for measuring coauthor diversity at an individual career level and comparing diversity between individual authors. For example, although H is a group-level indicator, the group can be adjusted to data as well as to research questions and purposes.

The direction of the relationship between the scientific productivity and H is not important *per se*. A negative correlation between the productivity and H also provides valuable information, when H is included in the research collaboration as a partial indicator. Actually, one might want to hypothesize a negative relationship between productivity and H and/or very low H if the organization in question has not been established to bring together researchers with similar interests. Besides, it is clear that a publication by several authors is not automatically better than a publication by a single author. On the basis of the present results, we propose that H should be used as a partial indicator of research collaboration based on coauthorships at the group-level, together with other complementary indicators. Research collaboration is not only between authors but also between organizations and countries, for instance.

As the ease of calculating CIs has been found to be a crucial element,[3] we used a simple index for our study. Simplicity was also the main reason that we used coauthorship as a reflection of research collaboration and the number of citations per publication as a reflection of scientific productivity. Despite criticism about coauthorship as an indicator of collaboration,[5,10] it has been widely used, and it is particularly suitable when the aim is to gather a large dataset using reasonable efforts. Nowadays, information about authors and their affiliations are, in most cases, easily available through the Internet without any additional costs. Due to the small dataset, the present findings do not allow us to make generalizations or draw strong conclusions concerning the effect of research collaboration on scientific productivity.

Financial Support and Sponsorship

Nil.

Conflicts of Interest

There are no conflicts of interest.

REFERENCES

1. Lee S, Bozeman B. The impact of research collaboration on scientific productivity. Soc Stud Sci 2005;35:678-702.
2. Savanur K, Srikanth R. Modified collaborative coefficient: A new measure for quantifying the degree of research collaboration. Scientometrics 2010;84:365-71.
3. Liao CH, Yen HR. Quantifying the degree of research collaboration: A comparative study of collaborative measures. J Informetr 2012;6:27-33.
4. Brew A, Boud D, Lucas L, Crawford K. Reflexive deliberation in international research collaboration: Minimizing risk and maximizing opportunity. High Educ 2013;66:93-104.
5. Bozeman B, Fay D, Slade CP. Research collaboration in universities and academic entrepreneurship: The-state-of-the-art. J Technol Transf 2013;38:1-67.
6. Rambur B. Creating collaboration: An exploration of multinational research partnerships. In: Brew A, Lucas L, editors. Academic Research and Researchers. 1st ed. Maidenhead, UK: Open University Press, McGraw-Hill Companies; 2009. p. 80-95.
7. Archer MS. Making Our Way through the World: Human Reflexivity and Social Mobility. 1st ed. New York: Cambridge University Press; 2007.
8. Nistor N, Daxecker I, Stanciu D, Diekamp O. Sense of community in academic communities of practice: Predictors and effects. High Educ 2015;69:257-73.
9. Bozeman B, Corley E. Scientist' collaboration strategies: Implications for scientific and technical human capital. Res Policy 2004;33:599-616.
10. Katz JS, Martin BR. What is research collaboration? Res Policy 1997;26:1-18.
11. Jacsó P. Errors of omission and their implications for computing scientometric measures in evaluating the publishing productivity and impact of countries. Online Inf Rev 2009;33:376-85.
12. Egghe L. Theory of collaboration and collaborative measures. Inf Process Manag 1991;27:177-202.
13. Rousseau R. Comments on the modified collaborative coefficient. Scientometrics 2011;87:171-4.
14. Li EY, Liao CH, Yen HR. Co-authorship networks and research impact: A social capital perspective. Res Policy 2013;42:1515-30.
15. Lawani SM. Quality, Collaboration and Citations in Cancer Research: A 268 Bibliometric Study. Ph.D. Dissertation, Florida State University; 1980.
16. Subramanyam K. Bibliometric studies of research collaboration: A review. J Inf Sci 1983;6:33-8.
17. Ajiferuke I. Burrell Q, Tague J. Collaborative coefficient: A single measure of the degree of collaboration in research. Scientometrics 1988;14:421-33.
18. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27:379-423.
19. Shannon CE. The Mathematical Theory of Communication. 1st ed. Urbana: University of Illinois Press; 1949.
20. Spellerberg IF, Fedor PJ. A tribute to Claude Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Wiener' index. Glob Ecol Biogeogr 2003;12:177-9.
21. Doane DP, Seward LE. Measuring skewness: A forgotten statistic? J Stat Educ 2011;19:1-18.