Extreme inequalities of citation counts in environmental sciences

Deepthi Chimalakonda[1], Alex R. Cook[2, 3], L. Roman Carrasco[1,*]

[1] Department of Biological Sciences, National University of Singapore, Singapore

[2] Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore.

[3] Yale-NUS College, National University of Singapore, Singapore.

Author for correspondence: dbsctlr@nus.edu.sg

Abstract

Well-established scientists are expected to be more likely to have their work recognised than early-career individuals and thus receive more citations. Estimating the degree of inequality in citation counts in environmental sciences can help identify the dynamics behind citation inequalities.

Using the scientific profiles of researchers in the Google Scholar database, we estimated the inequality in the distribution of citations in the disciplines of evolutionary biology, conservation biology and ecology. The data were modelled using short-tailed (exponential) and long-tailed power-law (Pareto) distributions. The inequality in performance in each distribution was assessed using Gini coefficients.

Citations counts per researcher presented Gini coefficients of 0.82–0.89, indicating extreme inequality. The results suggest that the reinforcement in citation counts due to seniority and previous success might be very strong. To produce meaningful comparisons of actual research impact using citation counts, factors such as lab size, collaborations or role in articles should ideally be controlled for.

Keywords: h-index; Matthew effect; preferential attachment; research merit; rich get richer; role-based h-index.

Introduction

Citation analysis is a bibliometric method increasingly used to assess the research output of universities, scientists, journals and even countries. Citations are used as basis for evaluating researchers for either positions or tenure, awarding grants and in determining the rank of universities [1-4] . Citation counts and the h-index are popular indicators of scientific merit [5,6]—where the h-index is the measure of the number of the researcher's articles that have at least *h* citations [1]. Although it has been found that the h-index is a better predictor of future achievement than citation numbers [7], citation counts still remain as one of the most commonly used measures of the research impact of individuals [8].

Citation counts may however not fully demonstrate the scientific merit of the researcher [9]: they often include self-citations and negative citations [10,11], might benefit from multiple-author publication practices [12,13] and they are not time-standardised to control for the fact that researchers who have been publishing for a long time tend to have more citations due to the additive effect resulting from the increase in the absolute number of published articles and the number of citations per article.

A second group of more intangible factors influencing citations are related to the previous success and seniority of the scientist. In other words, scientists who achieve fame are more likely to have their work recognised, for instance receiving more citations or collaborations, than early-career individuals producing similar quality of work [14,15]. This phenomenon has been termed "preferential attachment" or "the rich get richer effect". This means that resources are distributed proportionally to what an individual already has [16]. Other factors can also contribute to widen the gap between a few successful scientists and the rest: opportunities to collaborate with other top scientists in landmark articles, attainment of larger grants, having larger teams or attracting better quality students. These factors are especially prominent in environmental related sciences where principal investigators construct large groups and where multiple-author papers are common.

Preferential attachment has been widely studied in the field of economics, especially in connection to web links, wealth and population distributions [17]. Power laws tend to explain the pattern in these "rich get richer" models [18], where inequality is quite high. For instance the wealth distribution of the Forbes 400 list of US richest people follows a power law [19]. Similarly the income and wealth distributions in US and UK has been shown to follow a distribution with the high-end tail following a power law [20]. Power laws have also been used to show that the citations of academic articles are proportional to the number of citations that the article already has [18]. Such a process creates wider inequalities in the frequency distribution of the variable of interest. This inequality can be measured by using the Gini co-efficient, initially conceived to measure inequalities in incomes distributions. A Gini co-efficient of zero suggests perfect equality and one indicates maximum inequality [21].

Although the combined influence of reinforcement factors due to previous success is expected, it is very hard to estimate their influence on the inequality in citation counts. Here we aim to estimate the degree of inequality in citation counts for several environmental science disciplines using the recently available Google Scholar research profiles data.

Methods

Using the scientific profiles of researchers in the Google Scholar database, we extracted the citation counts for all scholars in various disciplines using the labels “ecology”, “conservation biology” and “evolutionary biology”.

The data were modelled using short-tailed (exponential) and long-tailed power-law (Pareto) distributions in the statistical environment R using the package VGAM [22]. A preliminary analysis indicated that Pareto type IV distributions obtained the best fit of all the Pareto types and was subsequently used for the analysis. The exponential (1) and Pareto type IV distributions (2) are expressed by the following formulas respectively:

$$F(y) = \lambda e^{-\lambda y} \quad (1)$$

$$F(y) = 1 - [1 + ((y - \mu) / \sigma)^{(1/\gamma)}]^{(-\alpha)} \quad (2)$$

Where λ is the rate, μ the location, σ the scale, α the shape and γ the inequality parameters respectively.

The models were fit to the data using Markov Chain Monte Carlo methods with a Gibbs sampler. The fit of the distributions was compared using the Akaike Information Criterion.

Pareto distributions can be used to characterize income distribution through their association with the Gini coefficient [21]. We employed the following formula to estimate the Gini coefficient of the Pareto IV functions [23]:

$$G = 1 - \left(\frac{\mu + 2\sigma\alpha B\left(2\alpha - \gamma, \gamma + 1\right)}{\mu + \sigma\alpha B(\alpha - \gamma, \gamma + 1)} \right),$$

where B denotes the Beta function.

Results

Pareto type IV provided the best fit for all the datasets (Table 1, Figure 1). Citations per researcher in Ecology, Evolutionary Biology and Conservation biology presented Gini coefficients of 0.89, 0.83 and 0.82 respectively, indicating a very high inequality. The inequality in citations in different fields of study was also obvious from our results where twenty percent of the scientists had more than eighty percent of all the citations (82% for ecology and conservation biology and 83% for evolutionary biology).

The citations within top cited environmental scientists present themselves high inequality. For instance, the most cited scientist in conservation biology has been cited 56,618 times and out of these 22,107 times correspond to a statistical book highly used for model analysis in the field. In a similar fashion, the top most cited scientist in evolutionary biology with 192,602 citations accrues a large share of his citations through the construction of widely used software for evolutionary genetics analysis.

Discussion

Our results show that there is extreme inequality in the citation counts in environmental sciences. For illustration, the greatest income inequality in the world occurs in South Africa, with a Gini coefficient of 0.7 [24] which is still below the 0.82–0.89 values obtained for citation counts distributions.

Extreme inequalities in environmental sciences could be increased due to factors related to the facility to collaborate with other successful scientists and to be invited as co-author in landmark articles or the attraction of the best students and postdoctoral researchers. Other institutional factors could be the promotion process or grant allocations. In addition to these, search engines such as Google Scholar and Web of Science tend to give higher order of appearance to results with high citations, creating a bias for researchers to read and cite these papers. At any rate, teasing out the actual research impact from the reinforcement in inequalities due to exogenous factors would need further research.

With views to produce more meaningful comparisons of scientific output, measures to alleviate the extreme inequalities in citation numbers would be needed. For instance to account for the facility to obtain collaborations and co-authorship of established scientists, variants of the h-index have been proposed: h-index methods that take into account the number of co-authors and relative position of the authors while assigning the rank could be employed [25]. Other role-based indices have also been suggested, where single author publications get higher ranking than multiple author ones or where only major contributions such as first and corresponding authors are considered for the estimation of the h-index [26,27]. These methods, however, would still not be able to account for institutional factors such as promotion, lab size, grant allocation and capacity to attract exceptional students.

Acknowledgements

D.C. and L.R.C. are thankful to the Singaporean MOE Tier 1 grant WBS R154000556112.

References

1. Hirsch JE (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: 16569.
2. Moed HF, Burger WJM, Frankfort JG, Van Raan AFJ (1985) The use of bibliometric data for the measurement of university research performance. Research Policy 14: 131-149.
3. Johnes J, Johnes G (1995) Research funding and performance in UK university departments of economics: a frontier analysis. Economics of Education Review 14: 301-314.
4. King DA (2004) The scientific impact of nations. Nature 430: 311-316.
5. King J (1987) A review of bibliometric and other science indicators and their role in research evaluation. Journal of information science 13: 261-276.
6. Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: Toward an objective measure of scientific impact. Proceedings of the National Academy of Sciences 105: 17268-17272.

7. Bornmann L, Daniel H-D (2005) Does the h-index for ranking of scientists really work? Scientometrics 65: 391-392.
8. Adam D (2002) Citation analysis: The counting house. Nature 415: 726-729.
9. Garfield E (1979) Is citation analysis a legitimate evaluation tool? Scientometrics 1: 359-375.
10. Fassoulaki A, Paraskeva A, Papilas K, Karabinis G (2000) Self-citations in six anaesthesia journals and their significance in determining the impact factor. British Journal of Anaesthesia 84: 266-269.
11. Fowler J, Aksnes D (2007) Does self-citation pay? Scientometrics 72: 427-437.
12. Persson O, Glänzel W, Danell R (2004) Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. Scientometrics 60: 421-432.
13. Hsu J-w, Huang D-w (2011) Correlation between impact and collaboration. Scientometrics 86: 317-324.
14. Merton RK (1968) The Matthew effect in science. Science 159: 56-63.
15. Eom Y-H, Fortunato S (2011) Characterizing and modeling citation dynamics. PloS one 6: e24926.
16. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286: 509-512.
17. Jeong H, Néda Z, Barabási A-L (2003) Measuring preferential attachment in evolving networks. EPL (Europhysics Letters) 61: 567.
18. Newman ME (2005) Power laws, Pareto distributions and Zipf's law. Contemporary physics 46: 323-351.
19. Levy M, Solomon S (1997) New evidence for the power-law distribution of wealth. Physica A: Statistical Mechanics and its Applications 242: 90-94.
20. Drăgulescu A, Yakovenko VM (2001) Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. Physica A: Statistical Mechanics and its Applications 299: 213-221.
21. Gini C (1921) Measurement of inequality of incomes. The Economic Journal 31: 124-126.
22. R Core Development Team (2010) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
23. Chotikapanich D (2008) Modeling income distributions and Lorenz curves [electronic resource]: Springer.
24. The World Bank (2011) South Africa Overview. The World Bank. Accessed on 12/9/2013 at: http://www.worldbank.org/en/country/southafrica/overview.
25. Wan J-k, Hua P-h, Rousseau R (2007) The pure h-index: calculating an author'sh-index by taking co-authors into account. COLLNET Journal of Scientometrics and Information Management 1: 1-5.
26. Hu X, Rousseau R, Chen J (2010) In those fields where multiple authorship is the rule, the h-index should be supplemented by role-based h-indices. Journal of Information Science 36: 73-85.
27. Schreiber M (2009) A case study of the modified Hirsch index hm accounting for multiple coauthors. Journal of the American Society for Information Science and Technology 60: 1274-1282.

Tables and figures captions

Table 1. AIC values for exponential and Pareto IV models fitted to the citation counts in environmental sciences.

Field	Model	AIC
Evolutionary Biology	Pareto IV	20075
	Exponential	21871
Ecology	Pareto IV	43531
	Exponential	47112
Conservation Biology	Pareto IV	33143
	Exponential	35879

Table 2. Parameter values and Gini coefficients of the Pareto IV models fitted to the data.

Field	Location	Scale	Inequality	Shape	Gini coefficient
Evolutionary Biology	0	6.79	1.43	1.78	0.89
Ecology	0	7.80	1.65	2.86	0.83
Conservation Biology	0	7.52	1.62	2.96	0.82

Figure 1. Comparison of Pareto IV and exponential fits to citations in ecology, conservation biology and evolutionary ecology.