Journal of Informetrics 11 (2017) 299–306

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Bibliometric author evaluation through linear regression on the coauthor network

Rasmus A.X. Persson

Department of Chemistry & Molecular Biology, University of Gothenburg, SE-412 96 Gothenburg, Sweden

ARTICLE INFO

Article history:
Received 9 November 2016
Received in revised form 15 January 2017
Accepted 16 January 2017

Keywords:
Multiple authorship
Statistical method
Coauthor contribution

ABSTRACT

The rising trend of coauthored academic works obscures the credit assignment that is the basis for decisions of funding and career advancements. In this paper, a simple model based on the assumption of an unvarying "author ability" is introduced. With this assumption, the weight of author contributions to a body of coauthored work can be statistically estimated. The method is tested on a set of some more than five-hundred authors in a coauthor network from the CiteSeerX database. The ranking obtained agrees fairly well with that given by total fractional citation counts for an author, but noticeable differences exist.

1. Introduction

Typical quantitative indicators of scientific productivity and quality that have been proposed—be it on the level of individuals, institutions or even whole geographic regions—are, in some form or another, ultimately based on the citation distribution to previous (and available) scientific works (in this paper referred to as "papers" for short for all types [books, regular articles, rapid communications, commentaries, proceedings, *etc.*]). A fairly extensive scientific literature exists on the subject of discriminating between individuals or scientific institutions, motivated to a large extent by the perceived need of the merit-based distribution of funding which is scarce in relation to the number of active scientists. Such indicators range from the simple (counting the number of papers and/or citations) to the more elaborate, such as the *h*-index (Bornmann & Daniel, 2005, 2007b; Bornmann, Mutz, & Daniel, 2008; Hirsch, 2005, 2007; Jin, 2006) and its many variants (Ausloos, 2015; Bras-Amorós, Domingo-Ferrer, & Torra, 2011; Egghe, 2006; Egghe & Rousseau, 2008; Jin, 2007; Jin, Liang, Rousseau, & Egghe, 2007; Kosmulski, 2006). For a recent and in-depth review of the fundamentals this topic (citation counting), see the paper by Waltman (2016). This comparison is in some schools of bibliometrics developed further in that the incoming citations to a paper are weighted by the importance of the citing source. This importance can be defined, for instance, from the number of citations the citing paper has itself received, or the number of citations of the citing author. For a review of this topic and an empirical investigation of its robustness, see the paper by Wang, Shen, and Cheng (2016).

In this paper, we are motivated by the confounding factor that coauthorship poses to any such analysis. Different options for dealing with this problem have been proposed. The simplest is to divide the credit equally among all contributing authors (Batista, Campiteli, Kinouchi, & Martinez, 2006; Schreiber, 2008) (known both as "fractional counting" or "normalized counting"); after that comes weighting author credit by a simple function of the author's position in the author list (Hagen, 2009; Sekercioglu, 2008; Zhang, 2009), or even more intricate schemes based on this notion (Aziz & Rozing, 2013). However, these alternatives cannot be motivated by more than "hunches" about how a particular "authorship culture" assigns credit. Clearly, a quantitative approach is more scientific than a qualitative, or worse, arbitrary one. Special mention is here given

E-mail address: rasmus.a.persson@gmail.com

http://dx.doi.org/10.1016/j.joi.2017.01.003
1751-1577/© 2017 Elsevier Ltd. All rights reserved.

to the papers by Tol (2011) and by Shen and Barabási (2014), in which intuitive statistical models are used to disentangle the coauthorship contributions.

Tol's (2011) idea may be summarized as follows. Whenever two authors write a joint paper and it is highly cited, the senior author of the pair[1] should receive a disproportionally large share of the citation credit. The rationale for this is that it is more typical of the senior author, judging from past experience, to write highly cited papers, and it is therefore reasonable to assume that her contribution is more responsible for the ultimate quality. With his method and a limited sample set comprising some fifty authors, Tol (2011) finds small deviations of up to 25% between his "Pareto weights" and what he terms "egalitarian weights" in which coauthorship credit is equally distributed.

Shen and Barabási (2014) agree with Tol (2011) on the principle of assigning more credit to the "senior author", but the algorithm to determine the actual credit assignment is different. To determine the "relative seniority" of each coauthor, their algorithm weighs both the number of papers by the author and the degree to which these papers share citations from papers citing the one under consideration. In this way, papers that are more "similar" to the one under consideration contribute more to the "seniority" of that coauthor when assigning the authorship credit.

The idea behind the present paper is basically the same, but the execution is different. Rather than assume a fixed form of a distribution like Tol (2011), we assume a fixed form for the underlying "ability" to produce said distribution in the first place. We then solve for this "author ability" statistically to find those authors who consistently manage to contribute to "high-quality" papers. Another difference, which also distinguishes the method from that by Shen and Barabási (2014), is that a junior author is not necessarily "punished" for publishing with a senior coauthor. If a paper is very successful compared to previous papers on the topic, it is not altogether unreasonable to assume that this atypical performance should be disproportionately credited to any authors not participating in the earlier work. However, in both Shen and Barabási (2014) and in Tol (2011), credit is instead disproportionately allocated to the senior author. Much like Tol (2011), the rigorous application of our method requires knowledge of complete coauthor networks, and can only be approximately applied otherwise. This is, however, more of a formal problem than a practical one.

2. Regression model for coauthorship contribution

We assume that the arbitrary author i has an unchanging ability a_i for contributing to scientific papers.[2] A paper α, once produced, possesses a "scientific quality" that we non-committally denote by q_α for now. This variable could be, for instance, the total number of citations or the rate of citation accumulation, to name a few. For notational simplicity, we define the elements, $f_{\alpha i}$, of a dimensionless "authorship tensor" $\mathbf{F}$, to be unity if author i contributes to paper α, and zero otherwise:

$$f_{\alpha i} = \begin{cases} 1, & \text{if } i \text{ is author of } \alpha \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

With these definitions, we now define a_i through,

$$\ln q_\alpha = \sum_{i=1}^{M_\mathrm{a}} f_{\alpha i} \ln a_i \tag{2}$$

where M_a is the total number of authors in the statistical sample, formally the number of individuals who have ever produced a work of science. In practical calculations, we limit ourselves to much smaller subsets of authors in a citation database. With modern computers, solving the complete system of equations is possible if one has access to the entire database. Typically, for individuals, the database is only partially accessible through search keywords of an online interface and the database in its entirety is not allowed (because of commercial contracts between the library and the database provider, for instance) to be downloaded and mined for its data. Such a limitation does not pose a greater problem than the reduction of the underlying statistical data.

Before we continue, we note that the choice of the logarithm function in Eq. (2) is judicious. First, it implies that "the whole is not equal to the sum of its parts" and is meant to capture at least some of the synergistic effects of a collaboration (as suggested, for instance, by Figg et al. (2006)): in other words, the relation between the number of authors and the resulting quality of the paper is taken to be non-linear rather than linear. Here, we follow Ke (2013) closely, but replace his "paper fitness" by our "author ability". Ke's model is more general, but we do not want to proliferate the number of fitting parameters needlessly. Second, since the value of q may vary over several orders of magnitude in typical cases (*vide infra*), the logarithm ensures a more modest range for the regression. This said, Eq. (2) is obviously an *Ansatz* chosen merely for its simple mathematical form rather than being based on some underlying physical understanding of research production within collaborations.

[1] Defined in terms of "Pareto weights" which are directly related to the average citations per article of an author.

[2] This assumption does not contradict the statement in Section 1 that "a senior author, judging from past experience," is more typically able to write highly cited papers. The senior author may always have been good at producing highly cited scientific output, but contrary to the case of the junior author, she has the credentials to back it up.