Journal of Informetrics 9 (2015) 800–808

Contents lists available at ScienceDirect

Journal of Informetrics

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

Quality versus quantity in scientific impact

Jasleen Kaur, Emilio Ferrara, Filippo Menczer, Alessandro Flammini, Filippo Radicchi*

Center for Complex Networks and Systems Research, School of Informatics and Computing, Indiana University, Bloomington, USA

ARTICLE INFO

Article history:
Received 26 January 2015
Received in revised form 14 July 2015
Accepted 14 July 2015
Available online 2 September 2015

Keywords:
Citation analysis
Scientific impact
Research evaluation
Quality
Quantity

ABSTRACT

Citation metrics are becoming pervasive in the quantitative evaluation of scholars, journals, and institutions. Hiring, promotion, and funding decisions increasingly rely on a variety of impact metrics that cannot disentangle quality from quantity of scientific output, and are biased by factors such as discipline and academic age. Biases affecting the evaluation of single papers are compounded when one aggregates citation-based metrics across an entire publication record. It is not trivial to compare the quality of two scholars that during their careers have published at different rates, in different disciplines, and in different periods of time. Here we evaluate a method based on the generation of a statistical baseline specifically tailored on the academic profile of each researcher. We demonstrate the effectiveness of the approach in decoupling the roles of quantity and quality of publications to explain how a certain level of impact is achieved. The method can be extended to simultaneously suppress any source of bias. As an illustration, we use it to capture the quality of the work of Nobel laureates irrespective of number of publications, academic age, and discipline, even when traditional metrics indicate low impact in absolute terms. The procedure is flexible enough to allow for the evaluation of, and fair comparison among, arbitrary collections of papers – scholar publication records, journals, and institutions; in fact, it extends a similar technique that was previously applied to the ranking of research units and countries in specific disciplines (Crespo, Ortuño-Ortí, & Ruiz-Castillo, 2012). We further apply the methodology to almost a million scholars and over six thousand journals to measure the impact that cannot be explained by the volume of publications alone.

1. Introduction

The interest in measuring scientific impact is no longer restricted to bibliometrics specialists, but extends to the entire scientific community. Many aspects of academic life are influenced by impact metrics: from the desire to publish in high-impact journals (Calcagno et al., 2012), to hiring, promotion and funding decisions (Bornmann & Daniel, 2006), and department or university rankings (Davis & Papanek, 1984; Liu & Cheng, 2005). Although the idea of measuring scientific impact is laudable, several fundamental aspects in the current evaluation methods are problematic; the use of existing citation-based metrics as proxies for "true" scientific quality of publications or scholars in practical contexts is often unsatisfactory (Adler, Ewing, & Taylor, 2009; Ke, Ferrara, Radicchi, & Flammini, 2015), or worse, misleading (Alberts, 2013; Editorial, 2005). Comparisons

* Corresponding author.
E-mail address: filiradi@indiana.edu (F. Radicchi).

http://dx.doi.org/10.1016/j.joi.2015.07.008
1751-1577/

among scholars, journals, and organizations are meaningful only if one takes into account the proper contextual information, such as discipline, academic age, publication and citation patterns.

Some of these issues can be addressed at the level of individual publications. Two important factors affecting the citations of an article are discipline and age. Once papers are divided into homogeneous sets according to these features, the populations within these classes can be used as baselines. One intuitive approach is that of assigning papers to citation percentiles (Leydesdorff, Bornmann, Mutz, & Opthof, 2011). Another possibility is to leverage the universality of citation distributions to measure relative citation counts (Radicchi & Castellano, 2012; Radicchi, Fortunato, & Castellano, 2008). The situation, however, becomes more challenging when we try to assess the quality of *aggregate* entities such as scholars, journals, or organizations. There have been several attempts to measure the impact of such entities that rely on aggregating across all the papers that can be attributed to the entity. Of course, the biases that affect the evaluation of individual papers are amplified when these aggregate measures are considered. Most impact metrics have been shown to be strongly biased by multiple factors when authors are considered (Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009; Duch et al., 2012; Kaur, Radicchi, & Menczer, 2013; Radicchi & Castellano, 2013) and corrections to mitigate biases due to discipline, multiple authors, and academic age have been proposed (Batista, Campiteli, & Kinouchi, 2006; Kaur et al., 2013; Schreiber, 2008; Sidiropoulos, Katsaros, & Manolopoulos, 2007; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). Unfortunately none of these corrections is effective against the whole spectrum of potential biases (Kaur et al., 2013).

The biases of impact metrics for researchers cannot be addressed with the same classification-based approach as for individual publications; scholars cannot be simply divided into categories that are simultaneously homogeneous for academic age and scientific discipline. First, it is not clear whether age should be quantified in terms of academic years of activity or total number of publications. Fixing only one of these two constraints would lead to a large variability for the other quantity. Accounting for both, instead, would produce sparsely populated categories of no practical use. Second, many researchers work on a range of different topics and in multiple disciplines (Albarrán, Crespo, Ortuño, & Ruiz-Castillo, 2011; Ruiz-Castillo & Costas, 2014; Sun, Kaur, Milojevic, Flammini, & Menczer, 2013), or change their research interests during their careers. Therefore, reducing a scholar's research to a restrictive scientific subject container makes little sense. Also here, focusing only on scholars who are involved in exactly the same set of topics would generate very sparse categories. The situation only worsens if one simultaneously takes into account age, disciplines, and their intricate longitudinal combinations. Here we adopt a strategy that addresses these issues by evaluating quality in the proper context.

2. Quality metrics

Our approach is similar to the one presented by Crespo and collaborators (Crespo et al., 2012). While the method can be applied to scholars, journals, institutions, or any aggregate set of papers, let us illustrate it in the case of a researcher. The idea is to generate a statistical baseline specifically tailored on the academic profile of the scholar; the term of comparison is not given by other individuals, but rather by artificial copies of that scholar. Each copy, or *clone*, has a publication record with identical publication years and subject categories as the researcher under observation. However, the citation profile is resampled: the number of citations of each paper is replaced by that of a paper randomly selected among those published in the same year and in the same discipline. The cloning procedure is illustrated in Fig. 1.

In essence, a clone encodes an academic trajectory that in number of papers, their publication years, and topics exactly corresponds to that of the scholar being cloned. One can compute any citation-based impact metric for a clone, given its citation profile. From a population of clones associated with a researcher profile, one can estimate the likelihood that the scholar's measured impact could be observed by chance, given her publication history. Since the publication history includes the number of publications, this procedure deals with the biases that affect this number, such as academic age and disciplinary publication practices. In other words, the procedure decouples quantity and quality, allowing to ask whether a certain level of impact can be explained by quantity alone, or an additional ingredient of *scientific quality* is necessary.

More specifically, consider a researcher r who published N_r papers, in specific years $\{y_1, y_2, \ldots, y_{N_r}\}$ and disciplines $\{s_1, s_2, \ldots, s_{N_r}\}$, that have received certain numbers of citations $\{c_1, c_2, \ldots, c_{N_r}\}$, where y_i, s_i and c_i indicate respectively the year of publication, the subject category, and the total number of citations accumulated by the i-th paper. Any citation-based impact metric for r can be calculated using this information, including simple ones, like total or average number of citations, or more sophisticated ones like the h-index (Hirsch, 2005). Let m_r be the observed score of the metric m for researcher r. A clone of r is generated by preserving the years and subject categories of the entire publication record of r, but replacing the number of citations c_i accumulated by any paper i with that of another paper randomly selected from the set of articles published in year y_i in subject category s_i. Once a clone is generated, we measure the value m'_r of the same impact metric m on its profile. After repeating this operation T times on as many independently generated clones, we compute the *quality score* q as the fraction of times that $m_r \geq m'_r$. We also compute the *standard score* $z_r = (m_r - \bar{m}_r)/\sigma_r$, where $\bar{m}_r$ and σ_r are the mean the standard deviation of m over the population of r's clones. Our numerical results are obtained using $T = 1000$.

2.1. Disciplines and publication venues

The cloning method relies on the classification of articles in subject categories. The discipline label s_i for a paper i may not be directly available in the data, but can be inferred by its publication venue v_i. Here, we use the term "publication venue" to refer to both scientific journals and conference proceedings. Mapping venues to disciplines and vice versa requires a Bayesian