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/© 2015 Elsevier Ltd. All rights reserved.

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

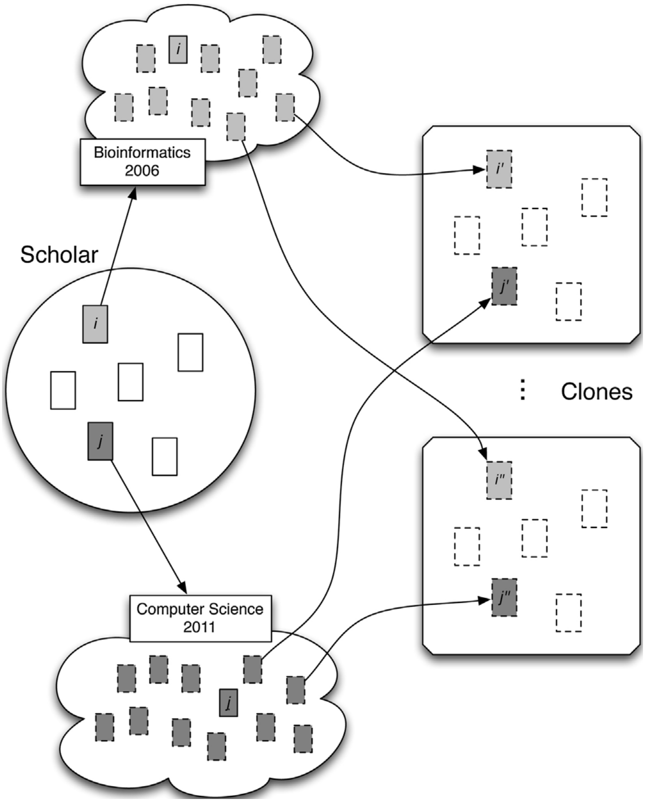

Fig. 1. Schematic illustration of the resampling technique used to generate the publication records of a scholar's clones. A scholar's paper i is replaced in a clone by a randomly selected paper i', published in the same year and in the same subject category. Similarly the paper is replaced by another paper i'', from the same set, in a different clone. The same resampling is applied to each paper for each clone.

framework to properly account for the many-to-many relationship between venues and subject categories. The rationale is that we want to preserve the relative sizes of disciplines to avoid under-sampling larger disciplines and oversampling smaller ones.

In practice, given a paper published in venue v in year y, we wish to replace its number of citations with those of another paper chosen at random among all publications in venue v' (that share at least one disciplinary category with v) and year y. To select v' we need to estimate the conditional probability $P_y(v'|v)$ that a paper published in year y and in venue v could have been potentially published in venue v'. Let us encode the classification of venues in subject categories via a matrix B, so that element $B_{vs} = 1$ if venue v is classified in subject category s, and $B_{vs} = 0$, otherwise. The probability that a randomly selected paper published in year y belongs to venue v is defined as

$$P_y(v) = \frac{N_y(v)}{\sum_u N_y(u)} \tag{1}$$

where $N_y(v)$ represents the total number of papers published in venue v in year y, and $\sum_u N_y(u)$ is the total number of papers published in year y. The probability that, given the venue v of publication, a paper belongs to category s is given by

$$P_y(s|v) = \frac{N_y(s)B_{vs}}{\sum_{s'} N_y(s')B_{vs'}} \tag{2}$$

where $N_y(s) = \sum_u B_{us} N_y(u)$ represents the total number of papers published in year y in venues belonging to category s. We can now formulate the probability that a randomly selected paper published in year y belongs to category s:

$$P_y(s) = \sum_v P_y(v) P_y(s|v). \tag{3}$$

We then use Bayes' theorem to calculate the probability that a paper in category s is published in venue v:

$$P_y(v|s) = \frac{P_y(s|v)P_y(v)}{P_y(s)}. \tag{4}$$

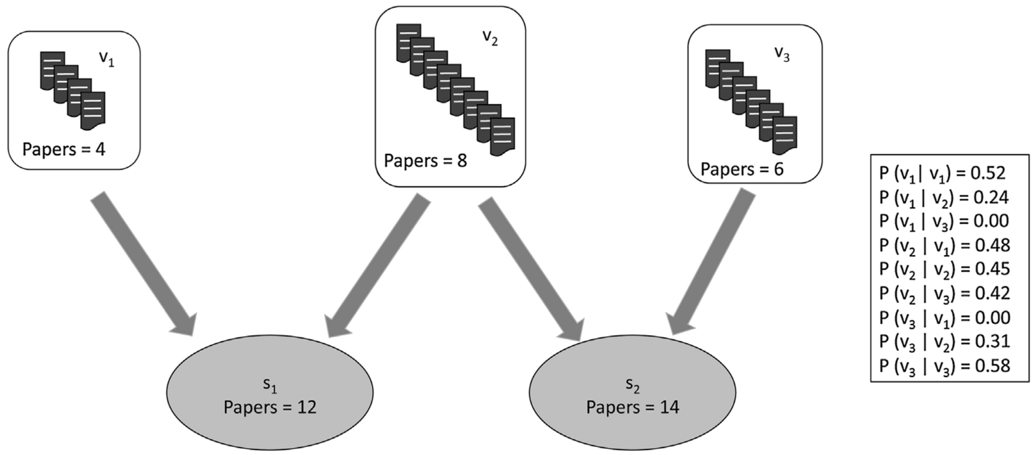

Fig. 2. Illustration of the Bayesian method for calculating resampling probabilities, in the case of three venues and two disciplines. A scholar's paper belonging to a venue is replaced by a randomly chosen paper from another venue, based on the probabilities shown in the box, and calculated as follows. Eq. (1) generates the values $P(v_1) = 2/9, P(v_2) = 4/9$, and $P(v_3) = 1/3$. The application of Eq. (2) produces $P(s_1|v_1) = 1, P(s_2|v_1) = 0, P(s_1|v_2) = 6/13, P(s_2|v_2) = 7/13, P(s_1|v_3) = 0$, and $P(s_2|v_3) = 1$. From Eq. (3) we have $P(s_1)=50/117$ and $P(s_2)=67/117$. Using Eq. (4) we calculate $P(v_1|s_1) = 13/25, P(v_1|s_2) = 0, P(v_2|s_1) = 12/25, P(v_2|s_2) = 28/67, P(v_3|s_1) = 0$, and $P(v_3|s_2) = 39/67$. Finally, Eq. (5) yields $P(v_1|v_1) = 13/25 = 0.52, P(v_1|v_2) = 6/25 = 0.24, P(v_2|v_1) = 12/25 = 0.48, P(v_2|v_2) = 9724/21775 \simeq 0.45, P(v_2|v_3) = 28/67 \simeq 0.42, P(v_3|v_2) = 21/67 \simeq 0.31$, and $P(v_3|v_3) = 39/67 \simeq 0.58$.

Finally, we have all the elements needed to compute

$$P_y(v'|v) = \sum_s P_y(v'|s) P_y(s|v). \tag{5}$$

Fig. 2 illustrates the resampling process captured by Eq. 5 in a simple example involving three venues and two subject categories. The Bayesian approach for the resampling of papers illustrated above represents the major difference between our method and the one introduced by Crespo et al. (Crespo et al., 2012). They considered a simpler strategy, in which every paper published in a venue belonging to more than one subject category is included with equal probability in all categories.

2.2. Bibliographic data

To implement our resampling procedure we consider all papers indexed in the Web of Science database published between 1970 and 2011 (Thomson Reuters, 2014a). These amount to 22,088,262 records classified as "articles," "proceedings papers," "notes," or "letters," and cover publications in 9696 science and social sciences venues. Based on the publication venue, we associate each article to one or more of 226 subject categories, as defined in the Journal Citation Reports database (Thomson Reuters, 2014b). The number of citations accumulated by each publication in our dataset was retrieved in March–April 2012. Authors are identified on the basis of first and middle name initials and full last name. Although we did not implement any disambiguation algorithm, we expect errors in author identification to account for less than 5% of the records (Radicchi, Fortunato, Markines, & Vespignani, 2009). We further restrict our attention only to authors with a publication record of at least 10 and at most 500 articles, filtering out many ambiguous names. The subsequent analysis is based on 996,288 author records matching these criteria.

3. Results

3.1. General properties of the quality score

Whereas the procedure described in Section 2 can be applied to any citation-based metric, let us first consider the h-index (Hirsch, 2005), widely used to represent the productivity and impact of researchers with a single number. In our analysis, we estimate the probability q that a clone's h is less than that of the corresponding scholar. Values of q close to 0 mean that the scholar's impact (as measured by h) is much smaller than one would expect from her publication profile (number of papers, and relative publication years and disciplines). Conversely, q close to 1 suggests that the author produced publications of consistently high quality.

Fig. 3 shows the relationship between the h value of four authors and those of their clones, yielding different values of q. In general, there is not a strict correspondence between the values of h and q. A high value of q is indicative of high quality even when the scholar's h is not high in absolute terms, as illustrated by the Fields medalist represented in Fig. 3A. Conversely, a high h does not necessarily imply high quality; most of the clones in Fig. 3D have higher h than that of the scholar. The distribution of h values for the clones is in general compatible with a bell-shaped distribution. This observation supports

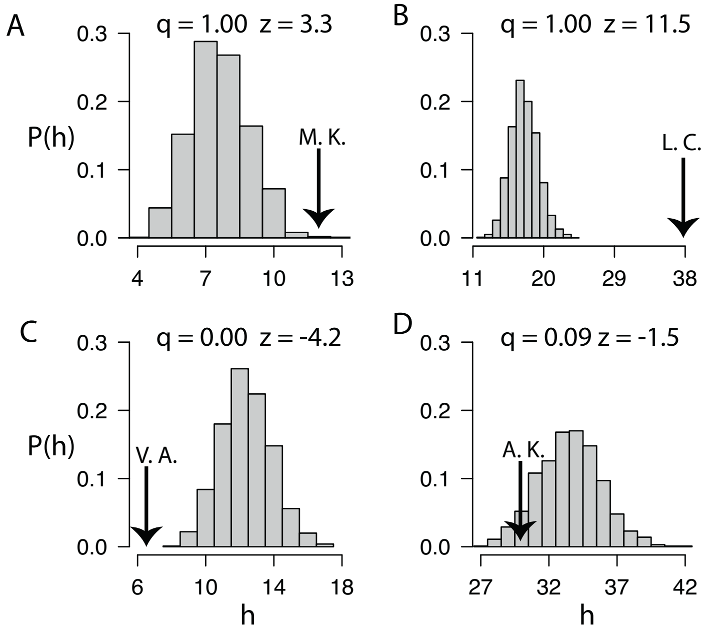

Fig. 3. Distributions of h for clones of four mathematicians, yielding different values of q and z. Arrows indicate actual h values. An author can have low h and high q (A), high h and high q (B), low h and low q (C), or high h and low q (D). The two scholars with high q are a Fields medalist and a Wolf Prize recipient.

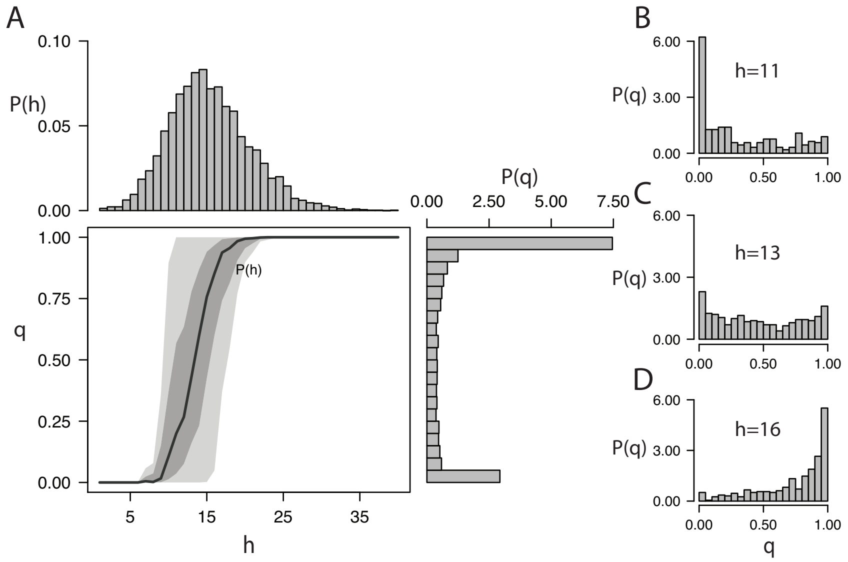

Fig. 4. Relationship between h and q for 4912 authors of exactly 50 papers. (A) Sharp transition of q around the critical value $h_c \simeq 13$. The black line represents the median q, and the gray areas represent the 50% and 95% confidence intervals. The distribution of h (top) and probability density function of q (right) are also shown. Plots (B–D) show the probability density functions of q values for $h = 11, 13, 16$.

our use of the standard score z as a related measure of scientific quality. The scores q and z convey similar information, but z provides higher resolution than q, especially for extremely low or high values of q. For example, the quality scores of the scholars of Fig. 3A and B are indistinguishable ($q = 1$) on the basis of the $T = 1000$ clones produced. Their standard scores, however, provide a basis for finer discrimination ($z = 3.3$ and $z = 11.5$, respectively).

General properties of the relation between h and q emerge when we consider the entire dataset. Fig. 4 shows that, as expected, $q \to 0$ and $q \to 1$ for small and large values of h, respectively. If we restrict ourselves to considering only authors with a fixed number of papers N, the transition between these two extremes is in general sharp and located at a critical value h_c that depends on N. The great majority of authors with $h < h_c$ have very low q, while most authors with $h > h_c$ have very high q. Overall, more than half of the researchers have extreme values of q: about 22.5% have $q \approx 0$, and 30% have $q \approx 1$.

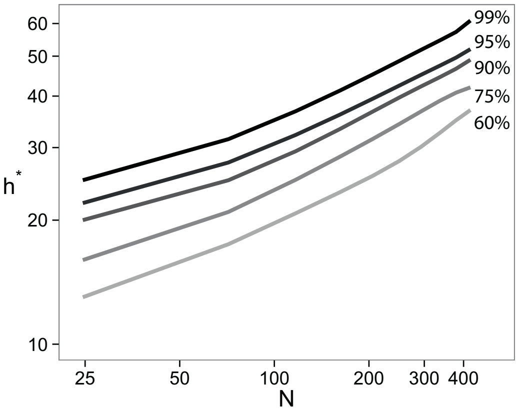

Fig. 5. Career phase diagram. Relationship between h^* and the number of papers N for 996,288 authors. Lines represent the values of h above which a fraction δ of authors in our dataset have q value larger than δ, for δ =0.60, 0.75, 0.90, 0.95, and 0.99.

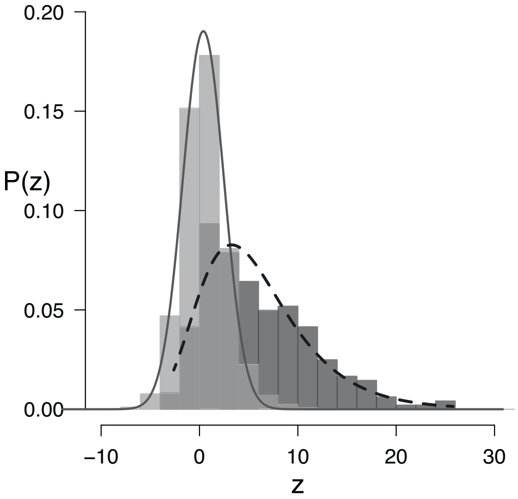

Fig. 6. Distribution of z for all scholars in our dataset (light gray). $P(z)$ is reasonably well fitted by a normal distribution with mean $\bar{z} = 0.4 \pm 0.1$ and standard deviation $\sigma_z = 2.0 \pm 0.1$ (continuous line). The distribution of the entire population is compared with the one of Nobel laureates (dark gray). The latter is compatible with a Gumbel distribution with parameters $\mu = 3.5 \pm 0.1$ and $\beta = 4.3 \pm 0.1$ (dashed line).

3.2. Critical lines of high quality

What is the value of h necessary to support a case of high scientific quality, given one's productivity (quantity)? The previous analysis suggests that such a value can be determined with some accuracy. Next we describe a procedure that can be employed to answer this question empirically. First, we bin authors according to the number of their published papers, N. For each bin, we then determine the value $h^*(N, \delta)$ defined by $P(q > 0.95 | h \geq h^*, N) = \delta$. $h^*(N, \delta)$ represents the value of h above which a fraction δ of scholars have a q value larger than 0.95. In Fig. 5 we draw $h^*(N, \delta)$ as a function of N for several values of δ. These phase diagrams separated by $h^*(N, \delta)$ can be interpreted as critical lines in the career trajectory of a scholar. We argue that impact characterized by $h \geq h^*(N, \delta = 0.95)$ provides strong evidence of high scientific quality.

3.3. Validating scientific quality

As already mentioned, q does not provide a sufficient resolution at the extremes; exceptionally good scholars, for example, all have $q \approx 1$. A similar problem was found for countries (Crespo et al., 2012). It is useful therefore to consider z scores. As Fig. 6 shows, $P(z)$ can be fitted relatively well by a normal distribution. Overall, about the 58% of the scholars have $z > 0$. Researchers with $z < -1$ and $z < -2$ amount to 24% and 11% of the population, respectively. Those with $z > 1$ and $z > 2$ represent 39% and 23% of the sample, respectively.

To test the ability of our quality score to recognize "true" scientific excellence (Albarrán & Ruiz-Castillo, 2012) we consider all Nobel laureates in the period 1970–2013. This set represents a small but ideal benchmark to check the validity of our method. The selection of Nobel recipients in fact does not depend on citations: laureates are identified each year by large

Table 1
Scientific quality of a sample of Nobel laureates. For each recipient we report the year of the award, the name of the laureate, the field of the award, and the h and z values associated with the academic profile.

Year	Laureate	Field	h-index	z-score
1991	P.-G. de Gennes	Physics	45	11.7
1991	B Sakmann	Medicine	103	20.0
1991	E. Neher	Medicine	87	15.2
1991	R.H. Coase	Economics	15	6.1
1998	D.C. Tsui	Physics	71	8.4
1998	R.B. Laughlin	Physics	51	5.1
1998	H.L. Storrmer	Physics	68	14.4
1998	J.A. Pople	Chemistry	109	23.7
1998	W. Kohn	Chemistry	42	5.7
1998	F. Murad	Medicine	74	10.7
1998	L.J. Ignarro	Medicine	75	10.0
1998	R. Furchgott	Medicine	17	3.4
2012	S. Haroche	Physics	56	13.0
2012	D.J. Wineland	Physics	68	14.4
2012	B.K. Kobilka	Chemistry	73	13.8
2012	J.B. Gurdon	Medicine	56	6.2
2012	L.S. Shapley	Economics	10	2.1

committees, often with the help of the entire scientific community. The distribution of z values for Nobel laureates is also shown in Fig. 6. Many Nobel recipients have very high z scores, reflecting the exceptional scientific quality revealed by their publication profiles. Their z distribution can be fitted well by a Gumbel distribution, which, interestingly, is often adopted to describe the statistics of extremes events (Gümbel, 1958). At the same time, we note that a few Nobel laureates have low z scores. In some cases this is the result of ambiguous names leading to profiles with inflated number of publications. For example, S.C.C. Ting (Physics, 1976) appears to have $z = -1.5$. Upon inspection we find that his publication record is composed of 655 publications, the majority of which are attributable to homonymous authors. As already noted for the general population of scientists, high values of z do not necessarily correspond to high h. In Table 1 we list a few examples of Nobel laureates in various disciplines and years. Among the majority of laureates with high h values, we find a few, such as R.H. Coase (Economics, 1991) and R. Furchgott (Medicine, 1998), with relatively low h values. In these cases, the z score is a more reliable indicator of exceptional scientific quality compared to h. We recognize that the metrics q and z of scientific quality, when applied to Nobel laureates, could be in principle affected by the boost in the number of citations typically observed after the award (Mazloumian, Eom, Helbing, Lozano, & Fortunato, 2011). Nevertheless, even when we focus on 2012 Nobel recipients only, we find consistently high z values, on occasion associated with relatively low h.

3.4. Analysis of publication venues

Although we focused on scholars until now, our general procedure is easily applicable to other aggregate entities, such as publication venues, with essentially no modification. Fig. 7 summarizes the results obtained for publication venues. We consider all publications in the period 1991–2000 and use h as impact metric for publication venues (Braun, Glänzel, & Schubert, 2006). We then apply our statistical procedure to calculate the critical $h^*(N, \delta)$ as described above. Specific examples of publication venues are marked in the diagram. Journals with large numbers of papers and high impact, such as *Nature*, *Science*, *PNAS*, *Cell*, and the *New England Journal of Medicine*, are well above the critical line $h^*(N, \delta = 0.95)$. High scientific quality can be achieved even if the number of papers in the publication venue is not large and the impact relatively small. This happens for example in the cases of the *Journal of Economic Literature* and *Nature Immunology*.

3.5. Applicability to different impact metrics

The proposed procedure can be used in conjunction with any arbitrary impact metric. As an illustration, let us consider the total number of citations c in place of the h-index. Fig. 8 plots the critical lines $c^*(N, \delta)$, defined analogously to $h^*(N, \delta)$, for authors and publication venues in the dataset. The z scores measured on the basis of the two impact metrics, h and c, are strongly correlated.

4. Discussion

The role of citation-based metrics in the quantitative evaluation of research activities has become so central that numbers derived from bibliographic data influence the behavior of scholars and other stakeholders on a daily basis (Garfield, 2006; King, 2004; Kinney, 2007). Although the use of citation-based metrics as proxies for “true” scientific impact is still debated (Adler et al., 2009), we believe that many controversial issues associated with current evaluation practices can be alleviated by designing better measure instruments. Measurements are meaningful only if taken in reference to proper terms of comparisons. Bibliometric numbers are instead often used as absolute quantities, and, as such, they do not convey much

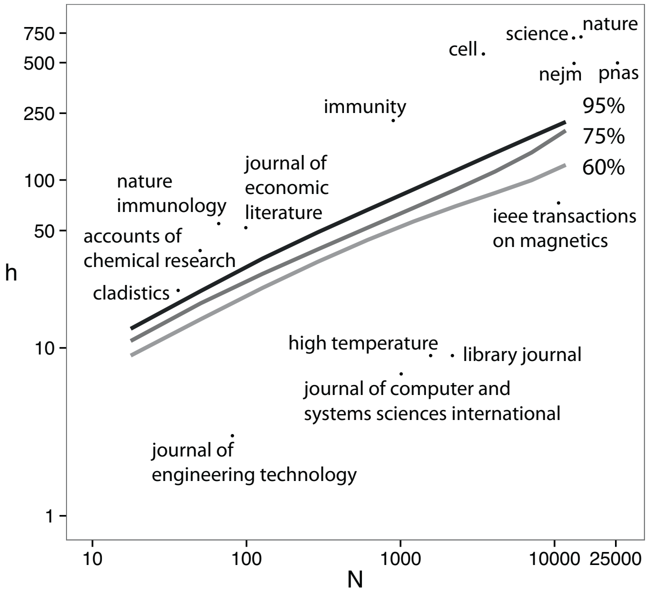

Fig. 7. Journal phase diagram. Relationship between h^* and N for 6129 journals. We consider only papers published in the period 1991–2000. The lines represent different values of δ. Several examples of journals are displayed.

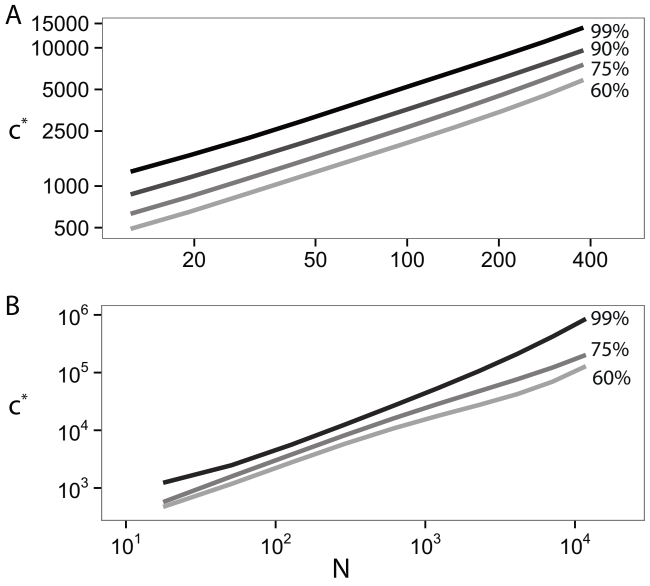

Fig. 8. Relationship between critical c^* and number of publications N for (A) 996,288 authors and (B) 6129 journals. Lines represent the values of c above which a fraction δ of authors or journals in our dataset have $q > 0.95$, for different values of δ.

information. While at the level of individual publications this issue can be easily addressed, at least when disciplinary biases are the concern (Leydesdorff et al., 2011; Radicchi et al., 2008), the proper evaluation of scholars represents a more challenging task (Kaur et al., 2013). Direct comparisons among individuals are not possible due to the intrinsic heterogeneity in publication records and career trajectories. Similar considerations are valid for other aggregate entities, such as publication venues, departments, and institutions, whose impact is generally quantified with additive metrics over sets of publications. Our proposal radically changes the point of view of the methodology currently in use to evaluate researchers and publication venues: the term of comparison is not given by other real entities, but artificial copies of the same entity. This procedure is very general, and allows us to assign a statistical significance to arbitrary impact metrics – from simple ones, like the total number of citations, to more complex ones, such as the h-index or the impact factor for publication venues. Furthermore, the procedure can be applied to compensate for any arbitrary source of bias, in place of or in addition to disciplines and academic age. For instance, one could create clones by preserving types of publications, say journals versus conference proceedings, or countries, languages, and so on. The only statistical requirement is the availability of representative sets of publications in each category.

We studied a large set of scientific publications to show the utility of our approach in assessing scientific quality irrespective of number of publications and impact. We demonstrated that the procedure is capable of singling out exceptional journals and scholars. Others have shown that a similar procedure can be applied to perform the same quality analysis for even more heterogeneous entities, such as research groups, institutions, and countries (Crespo et al., 2012). Our proposed use of z scores overcomes the resolution problem of q reported in those cases.

An additional merit of the quality evaluation system presented here is to promote parsimonious publishing strategies: increasing the number of publications also increases the critical threshold of impact necessary to demonstrate scientific quality. This will hopefully discourage questionable practices such as “salami publishing,” “self-plagiarism,” and the “minimum publishable unit” (Broad, 1981) that are incentivized by quantity-based impact metrics.

Acknowledgements

We are grateful to anonymous referees for constructive criticism and especially for bringing to our attention the closely related work by Crespo et al. (Crespo et al., 2012). We also acknowledge helpful discussions with current and past members of the Center for Complex Networks and System Research at Indiana University (`cnets.indiana.edu`). This work is supported in part by NSF (grants SMA-1446078 and CCF-1101743) and the J.S. McDonnell Foundation (grant 220020274).

References

Adler, R., Ewing, J., & Taylor, P. (2009). Citation statistics. *Statistical Science*, *24*(1), 1.

Albarrán, P., & Ruiz-Castillo, J. (2012). The measurement of scientific excellence around the world. In *Working Paper, Economic Series 12-08* Universidad Carlos III de Madrid, Departamento de Economía.

Albarrán, P., Crespo, J. A., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, *88*(2), 385–397.

Alberts, B. (2013). Impact factor distortions. *Science*, *340*(6134), 787. http://dx.doi.org/10.1126/science.1240319

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). h-index: A review focused in its variants computation and standardization for different scientific fields. *Journal of Informetrics*, *3*(4), 273–289.

Batista, P. D., Campiteli, M. G., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, *68*(1), 179–189.

Bornmann, L., & Daniel, H.-D. (2006). Selecting scientific excellence through committee peer review – a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, *68*(3), 427–440.

Braun, T., Glänzel, W., & Schubert, A. (2006). A hirsch-type index for journals. *Scientometrics*, *69*(1), 169–173.

Broad, W. (1981). The publishing game: Getting more for less. *Science*, *211*(4487), 1137–1139. http://dx.doi.org/10.1126/science.7008199

Calcagno, V., Demoinet, E., Gollner, K., Guidi, L., Ruths, D., & De Mazancourt, C. (2012). Flows of research manuscripts among scientific journals reveal hidden submission patterns. *Science*, *338*(6110), 1065–1069.

Crespo, J. A., Ortuño-Ortí, I., & Ruiz-Castillo, J. (2012). The citation merit of scientific publications. *PLOS ONE*, *7*(11), e49156.

Davis, P., & Papanek, G. F. (1984). Faculty ratings of major economics departments by citations. *The American Economic Review*, 225–230.

Duch, J., Zeng, X. H. T., Sales-Pardo, M., Radicchi, F., Otis, S., Woodruff, T. K., et al. (2012). The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLOS ONE*, *7*(12), e51332.

Editorial. (2005). Not-so-deep impact. *Nature*, *435*(7045), 1003–7431. http://dx.doi.org/10.1038/4351003b

Gümbel, E. J. (1958). *Statistics of extremes*. Columbia Univ. Press.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, *295*(1), 90–93.

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), 16569–16572.

Kaur, J., Radicchi, F., & Menczer, F. (2013). Universality of scholarly impact metrics. *Journal of Informetrics*, *7*(4), 924–932. http://dx.doi.org/10.1016/j.joi.2013.09.002

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences U.S.A.*, *112*, 7426–7431.

King, D. A. (2004). The scientific impact of nations. *Nature*, *430*(6997), 311–316.

Kinney, A. (2007). National scientific facilities and their science impact on nonbiomedical research. *Proceedings of the National Academy of Sciences*, *104*(46), 17943–17947.

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables on citation analysis one more time: Principles for comparing sets of documents. *Journal of the American Society for Information Science and Technology*, *62*(7), 1370–1381.

Liu, N. C., & Cheng, Y. (2005). The academic ranking of world universities. *Higher Education in Europe*, *30*(2), 127–136.

Mazloumian, A., Eom, Y.-H., Helbing, D., Lozano, S., & Fortunato, S. (2011). How citation boosts promote scientific paradigm shifts and nobel prizes. *PLoS ONE*, *6*(5), e18975.

Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLOS ONE*, *7*(3), e33833.

Radicchi, F., & Castellano, C. (2013). Analysis of bibliometric indicators for individual scholars in a large data set. *Scientometrics*, *97*(3), 627–637.

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*(45), 17268–17272.

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, *80*(5), 056103.

Ruiz-Castillo, J., & Costas, R. (2014). The skewness of scientific productivity. *Journal of Informetrics*, *8*(4), 917–934.

Schreiber, M. (2008). To share the fame in a fair way, hm modifies h for multi-authored manuscripts. *New Journal of Physics*, *10*, 040201.

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, *72*(2), 253–280.

Sun, X., Kaur, J., Milojevic, S., Flammini, A., & Menczer, F. (2013). Social dynamics of science. *Scientific Reports*, *3*, 1069. http://dx.doi.org/10.1038/srep01069

Thomson Reuters. (2014a]). *Web of science*. http://wokinfo.com

Thomson Reuters. (2014b]). *Journal citation reports*. http://thomsonreuters.com/journal-citation-reports

Waltman, L., van Eck, N., van Leeuwen, T., Visser, M., & van Raan, A. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47.