Universality of scholarly impact metrics

Jasleen Kaur\textsuperscript{1}, Filippo Radicchi\textsuperscript{2}, Filippo Menczer\textsuperscript{1}

\textsuperscript{1} Center for Complex Networks and Systems Research  
School of Informatics and Computing, Indiana University, Bloomington, USA  
\textsuperscript{2} Departament d’Enginyeria Quimica, Universitat Rovira i Virgili, Tarragona, Spain

Abstract

We present a method to quantify the disciplinary bias of any scholarly impact metric, and introduce a simple universal metric that allows to compare the impact of scholars across scientific disciplines.

Objective evaluation of scientific production — its quantity, quality, and impact — is quickly becoming one of the central challenges of science policy with the proliferation of academic publications and diversification of publishing outlets \cite{1}. Many impact metrics have been and continue to be proposed \cite{2}, most of them based on increasingly sophisticated citation analysis \cite{2} \cite{3}. These metrics have found wide applicability in the evaluation of scholars, journals, institutions, and countries \cite{4,5,6,7}. Unfortunately, there is very little work on quantitative assessment of the effectiveness of these metrics \cite{8,9} and the few existing efforts are proving highly controversial \cite{10}. This is alarming, given the increasingly crucial role of impact analysis in grant evaluation, hiring, and tenure decisions \cite{11}.

Discipline bias is probably the most critical and debated issue in impact metric evaluation. Publication and citation patterns vary wildly across disciplines, due to differences in breadth and practices. These differences introduce strong biases in impact measures — a top scholar in biology has a very different publication and citation profile than one in mathematics. This has led to a recent burst of interest in \textit{field normalization} of impact metrics, and the emergence of many “universal” metrics that claim to compensate for discipline bias \cite{12}. Fig. 1(a) illustrates the idea of field normalization. If we rank scholars across all disciplines according to an unbiased (universal) metric, a scholar in the top 5\% among mathematicians should be ranked the same as a scholar in the top 5\% among biochemists. A biased
metric on the other hand may favor some disciplines and penalize others.

An objective, quantitative assessment of metric universality is missing to date. To fill this void, we introduce a *universality index* to evaluate and compare the bias of different metrics. Our index allows for the first time to gauge a metric’s capability to compare the impact of scholars across disciplinary boundaries, creating an opportunity for, say, mathematicians and biologists to be evaluated consistently.

The proposed universality index looks at how top authors according to a particular metric are allocated across different disciplines, and compares this distribution with one obtained from a random sampling process. This approach is inspired by a method for comparing expected and observed proportions of top cited papers to evaluate normalized citation counts [13]. The idea is that each discipline should be equally represented in a sample of top authors. For example, if we rank scholars across all disciplines according to an unbiased (universal) metric, the top 5% of scholars should include the top 5% of mathematicians, the top 5% of biologists, and so on. In other words, the percentage of top scholars in each discipline should not depend on the size of the discipline. Of course the number of scholars in each discipline should be proportional to the size of that discipline.

Suppose each author is assigned a discipline $d$. Selecting a fraction $z$ of top scholars from the entire set according to a universal metric should be equivalent to sampling a fraction $z$ of scholars at random, where the expected fraction of scholars selected from each discipline $d$ is $E[f_{z,d}] = z$. Such a random process provides us with a null model for the values of $f_{z,1}, f_{z,2}, \ldots, f_{z,D}$, observed in the empirical data for all $D$ disciplines and calculated according to a particular metric $m$. To obtain a quantitative criterion for the universality of $m$ with respect to a set of $D$ disciplines and a fraction $z$ of top scholars, we compute the *universality* of metric $m$ as

$$u_m(z) = 1 - \frac{1}{D} \sum_{d=1}^{D} \left| \frac{f_{z,d}^m}{z} - 1 \right|^\alpha$$

where the parameter $\alpha$ tunes the relative importance given to small versus large deviations from the expected fractions. Here, we use $\alpha = 1$. If $u_m(z)$ is high (close to one), the proportion of top scholars from each discipline is close to $z$, and therefore the impact measure $m$ is able to compensate for discipline
bias. This definition of universality satisfies the basic intuition that all metrics are unbiased in the limit $z = 1$. To eliminate the dependence of the universality assessment on a particular selectivity $z$, we can finally define the \textit{universality index} $\bar{u}_m = \int_0^1 u_m(z)dz$. Further details on the null model and universality definitions can be found in appendix.

To illustrate the usefulness of our index, let us analyze the universality for a set of well-known impact measures across a set of scholarly disciplines. We consider 12 disciplines from the Thomson-Reuters JCR classification (see appendix). Some of the metrics under consideration, such as the $h$ index, are widely adopted \cite{14}. Others, such as the new crown indicator \cite{15} and Batista’s metric \cite{16}, are designed with the explicit goal of field normalization. We additionally propose a new metric $h_s = h/\langle h \rangle_d$ that simply normalizes $h$ by the average across authors in the same discipline. See appendix for further details on all of the metrics. As evident in Fig. 1(b), some metrics are more universal than others; in other words, when we select the top 5\% of all scholars, we find close to 5\% of scholars from each of the considered disciplines, consistently with the null model (grey area). Fig. 1(c) shows that according to $u(5\%)$, two of the metrics appear to be least biased: Batista’s $h_{i,norm}$ and our own $h_s$. These are consistent with the unbiased model at $z = 5\%$, while the other metrics are not. Fig. 1(d) shows how the universality of each metric depends on the selectivity $z$. As we select more top scholars, the bias of all metrics decreases; $u(z) \to 1$ as $z \to 1$ by definition. For selectivity $z < 40\%$, the two best metrics display high universality, as illustrated by the overlap of the corresponding curves with the expectation of the null model (grey area). Table 1(e) reports the values of the universality index $\bar{u}$ integrated across $z$. The fluctuations of the null model have standard deviation $\sigma = 0.005$, therefore we do not consider differences in the third decimal digit of the universality index $\bar{u}$ significant; values are rounded to the second decimal digit in the table, and the differences shown are significant. According to this summary, $h_{i,norm}$ and $h_s$ are the most universal among the impact metrics considered. Their universality indices $\bar{u} = 0.94$ are statistically equivalent to each other. The computation of $h_s$ is however much simpler, as it does not require co-author metadata. While our definition of universality assumes that authors are associated with disciplines, the

3
results of our analysis are not dependent on the JCR classification; similar results are obtained with alternative classifications of disciplines or larger values of the exponent $\alpha$ (see appendix).

While discipline bias is quickly being recognized as a key challenge for objective assessment of impact, it is difficult to evaluate the claims of universality for the multitude of proposed metrics. The index presented here is the first quantitative gauge of universality that can be readily applied to any existing metric. The present analysis points to $h_s$ as an impact metric that is intuitive, easy to compute, and universal.

**Acknowledgements**

Thanks to Santo Fortunato, Alessandro Flammini, Claudio Castellano, and Yong-Yeol Ahn for helpful feedback on earlier versions of this manuscript. Xiaoling Sun, Diep Thi Hoang, Mohsen JafariAsbagh, and Lino Possamai provided technical support. We acknowledge financial support from the IU School of informatics and Computing, the Lilly Endowment, and NSF (award IIS-0811994) for funding the computing infrastructure that hosts the Scholarometer service.
Figure 1: (a) Effect of field normalization on ranking bias. We rank the top 5% of authors in 7 JCR disciplines according to two metrics, $h$ and $h_s$ (see text). We compare the rank (top percentile) globally across disciplines versus locally within an author’s own discipline. Due to discipline bias, biochemists are favored and mathematicians are penalized according to $h$, as illustrated by the two highlighted authors. The global ranking according to the normalized metric $h_s$ is more consistent with the rankings within disciplines. (b) Illustration of discipline bias for impact metrics $h$ and $h_s$. The analysis is based on empirical data from the Scholarometer system, which provides discipline annotations for scholars and associated citation material [17, 18]. We consider 7 JCR disciplines spanning science and social sciences (see appendix). Across these disciplines, we select the top 5% of authors according to each metric. We then measure the percentage of authors from this selection that belong to each discipline. The $h$ index favors science disciplines and penalizes some social sciences and mathematics. In this and the following plots, grey areas represent 90% confidence interval of unbiased samples, calculated according to a hypergeometric distribution that accounts for the finite size of the sample [19] (see appendix). (c) Universality $u(z)$ for ten impact metrics and selectivity $z = 5\%$. (d) Universality $u(z)$ as a function of selectivity $z$. (e) Universality index $\bar{u}$ for the ten metrics obtained by integrating the curves in (d) across values of $z \in (0, 1)$. 
Appendix

1 Null Model

Consider a set of $N$ authors in $D$ categories. For simplicity, let us assume that each author belongs to only one category. Let $N_d$ be the number of authors in category $d$. Each author has a score calculated according to the rules of the particular indicator we want to test. Imagine extracting the top fraction $z$ of authors according to their scores. This list has $n_z = \lfloor zN \rfloor$ authors. If the numerical indicator is fair, the selection of an author in category $d$ should depend only on the category size $N_d$, and not on other features that may favor or hinder that particular category. Under these conditions, the number of authors $m^*_d$ in category $d$ that are part of the top $z$ is a random variate obeying the hypergeometric distribution:

$$P(m^*_d|n_z, N, N_d) = \binom{N_d}{m^*_d} \binom{N - N_d}{n_z - m^*_d} / \binom{N}{n_z}$$  \hspace{1cm} (1)

where $\binom{x}{y} = \frac{y!}{x!(x-y)!}$ is a binomial coefficient that calculates the total number of ways in which $y$ elements can be extracted out of $x$ total elements. Eq. 1 describes a simple urn model [20], where elements (authors in our case) are randomly extracted from the urn without replacement. With this statistical model we can calculate the expected number of authors in category $d$ present in the top fraction $z$ as $E(m^*_d) = n_z N_d / N$. Moreover, we can make use of Eq. 1 for estimating confidence intervals or other relevant statistical quantities. The process leading to Eq. 1 is simulated $10^3$ times to obtain the grey areas in Fig. 1(b,c,d) in the main text, and in Figs. A2 and A3.

2 Universality Test

The universality index is defined as

$$\bar{u}_m = \int_0^1 u_m(z) dz,$$
where \( u_m(z) \) is the universality score for a measure \( m \) and selectivity \( z \). We numerically approximate the integral as:

\[
\bar{u}_m \simeq \sum_{q=1}^{Q} u_m(q \cdot dq) dq,
\]

where we set \( dq = 0.01 \) and \( Q = 99 \). The universality score \( u_m(z) \) is calculated as

\[
u_m(z) = 1 - \frac{1}{D} \sum_{d=1}^{D} \left| \frac{f_{z,d}}{z} - 1 \right|^\alpha.
\]

where \( \alpha \) is a parameter (see below), \( D \) is the total number of disciplines in which authors are classified, and \( f_{z,d}^m \) is the fraction of authors from discipline \( d \) ranked by metric \( m \) in the top \( z \). Note that \( u_m(z) \leq 1 \); it can take negative values in contrived biased scenarios. An alternative definition would normalize the deviations from the expected fractions by the variance within each discipline, however this approach would have decreasing universality as \( z \to 1 \) due to the increasing variance. This would violate our basic intuition that all metrics are unbiased in the limit \( z = 1 \).

To evaluate the statistical significance of differences in values of the universality index \( \bar{u} \) for different metrics, we need to estimate the fluctuations of this measure. Let us consider the variations in the values of \( \bar{u}_{\text{null}} \) obtained by simulating the null model for \( z \in (0, 1) \). Running \( 10^3 \) simulations yields a standard deviation \( \sigma_{\text{null}} = 0.005 \). We therefore round \( \bar{u} \) values to the second decimal digit.

## 3 Data

We used the data collected by Scholarometer (scholarometer.indiana.edu) from November 2009 to August 2012. Scholarometer is a social tool for scholarly services developed at Indiana University, with the goal of exploring the crowdsourcing approach for disciplinary annotations and cross-disciplinary impact metrics [17][18]. Users provide discipline annotations (tags) for queried authors, which in turn are used to compare author impact across disciplinary boundaries. The data collected by Scholarometer is available via an open API. We use this data to compute several impact metrics for authors belonging to various disciplines, and test the universality of these metrics. At the time of writing, the database
has collected citation data about 38 thousand authors of 2.2 million articles in 1,300 disciplines. Further
statistics for authors and disciplines are available on the Scholarometer website [18].

4 Impact Measures

The bibliometrics literature contains a plethora of scholarly impact metrics, and it is not feasible to
evaluate all of them. Therefore we focus on a small set of metrics that are widely adopted and/or
specifically designed to mitigate discipline bias. Our analysis of universality is performed on the following
impact metrics:

c\textsubscript{avg} is the average number of citations received by an author’s articles.

h index is defined as the maximum number of articles h such that each has received at least h cita-
tions [14]. The h index is the most widely adopted impact metric. It summarizes the impact of a
scholar’s career using a single number without any threshold.

Redner’s index $c_{total}^{1/2}$ is defined as the square root of the total number of citations received by the
articles of an author [21].

h\textsubscript{m} index attempts to apportion citations fairly for papers with multiple authors [22]. It counts the
papers fractionally according to the number of authors. This yields an effective rank, which is
utilized to define $h_m$ as the maximum effective number of papers that have been cited $h_m$ or more
times.

g index is the highest number g of papers that together receive $g^2$ or more citations [23]. It attempts to
mitigate the insensitivity of the h index to the number of citations received by highly cited papers.

$i_{10}$ is proposed by Google and is defined as the number of articles with at least ten citations each [24].

h\textsubscript{f} index was proposed as a universal variant of h [13]. The number of citations c received by each
paper is normalized by the average number of citations $c_0$ for papers published in the same year
and discipline. The rank of each paper $n$ is rescaled by the average number $n_0$ of papers per author written in the same year and discipline. The $h_f$ index of the author is the maximum rescaled rank $h_f$ such that each of the top $h_f$ papers has at least $h_f$ rescaled citations.

Batista’s $h_{i,norm}$ involves normalizing the total number of citations in the $h$-core (the papers that contribute to the $h$ index) by the total number of authors contributing to them. The resulting $h_i$ of each author is then normalized by the average $h_i$ of the author’s discipline [16].

New crown indicator $(c/c_0)_{avg}$ was proposed by Lundberg [15] as the item oriented field-normalized citation score (FNCS) and implemented by Waltman et al. [25]. It is a modification of the crown indicator [26], calculated as the average field-normalized number of citations $c/c_0$ across an author’s publications.

$h_s$ index is proposed here as a normalization of the $h$ index by the average $h$ of the authors in the same discipline. Numerical tests show that the distribution of $h$ is not scale-free and therefore the mean is well defined. Despite its simplicity, we are not aware of this metric being previously defined in the literature. Note that within a discipline, $h_s$ produces the same ranking as $h$. Therefore, $h_s$ is very similar to the percentile score but slightly easier to compute. Percentiles have been proposed for normalization of journal impact factors [27].

5 Sensitivity Analysis

Here we test the robustness of our findings with respect to several variations of our method, namely different ways to classify authors into disciplines, different selectivity values, and different exponents in the definition of universality.

5.1 Discipline Definitions

To test the universality of the different impact metrics, we consider three distinct ways to define disciplines, i.e., to sample authors from multiple disciplines. When a user queries the Scholarometer system,
she has to annotate the queried author with at least one discipline tag from the JCR science, social sciences, or arts & humanities indices. Additionally, the user may tag the author with any number of arbitrary (JCR or user-defined) discipline labels. Based on these annotations, we consider three disciplinary groupings of authors:

**ISI:** The 12 JCR disciplines with the most authors (see Table A1). Results based on this method are presented in the main text.

**User:** The top 10 user-defined disciplines (Table A2).

**Manual:** 11 manually constructed groups of related disciplines (Table A3).

Table A1: Top JCR (ISI) disciplines.

| Discipline                                      | Authors | $\langle h \rangle$ |
|------------------------------------------------|---------|---------------------|
| 1. computer science, artificial intelligence   | 1,922   | 15.96               |
| 2. biology                                    | 1,147   | 19.66               |
| 3. economics                                  | 972     | 17.02               |
| 4. engineering, electrical & electronic        | 936     | 14.77               |
| 5. neurosciences                              | 840     | 22.95               |
| 6. political science                          | 794     | 15.81               |
| 7. psychology                                 | 774     | 21.18               |
| 8. biochemistry & molecular biology           | 766     | 22.37               |
| 9. sociology                                  | 749     | 16.70               |
| 10. mathematics                               | 516     | 13.55               |
| 11. philosophy                                | 501     | 13.63               |
| 12. information science & library science      | 480     | 11.15               |

Here we extend the analysis in the main paper to the two additional categorization. Fig. A2 reproduces Fig. 1(b) in the main text, and extends it for more JCR disciplines and the two additional discipline definitions (User and Manual). The results in all cases are similar. Fig. A3, and the tables therein, replicate Fig. 1(d) and Table 1(e) in the main text for the two additional categorizations. With a few exceptions, the ranking of impact metrics is consistent across categorizations. In all cases, $h_s$ and $h_{i,norm}$ are the most universal (least biased) metrics.
Figure A2: Percentage of authors belonging to different disciplines according to ISI JCR (top), manually-clustered (middle), and user-defined (bottom) disciplines, who are ranked by each metric in the top z = 5%, 20%. Gray areas bound the 90% confidence intervals obtained from the null model.
Figure A3: Universality $u(z)$ versus selectivity $z$ and universality index $\bar{u}$ of various metrics for ISI JCR (top), manually-clustered (middle), and user-defined (bottom) disciplines. The results are not particularly sensitive to different values of the exponent ($\alpha = 1$ on left and $\alpha = 2$ on right). Gray areas in the figure display 90% confidence intervals computed through the null model.
5.2 Selectivity

For each discipline categorization and impact metric, we obtained the actual distribution of authors across disciplines for each value of selectivity \( z \). Fig. A2 reproduces Fig. 1(b) in the main text, and extends it for a larger value of \( z \). The results in all cases are similar.

5.3 Exponent in Definition of Universality

Fig. A3, and the tables therein, replicate Fig. 1(d) and Table 1(e) in the main text for different values of the exponent \( \alpha \). With a few exceptions, the ranking of impact metrics is consistent for different exponents.

References

[1] Abbott, A. et al. Do metrics matter? Nature 465, 860–862 (2010).
[2] Van Noorden, R. Metrics: A profusion of measures. Nature 465, 864–866 (2010).
[3] Lane, J. Let’s make science metrics more scientific. Nature 464, 488–489 (2010).
[4] Garfield, E. The history and meaning of the journal impact factor. JAMA: the journal of the American Medical Association 295, 90–93 (2006).
[5] Davis, P. & Papanek, G. Faculty ratings of major economics departments by citations. The American Economic Review 74, 225–230 (1984).
[6] Kinney, A. National scientific facilities and their science impact on nonbiomedical research. Proceedings of the National Academy of Sciences 104, 17943–17947 (2007).
[7] King, D. The scientific impact of nations. Nature 430, 311–316 (2004).
[8] Bollen, J., Van de Sompel, H., Hagberg, A. & Chute, R. A principal component analysis of 39 scientific impact measures. PloS one 4, e6022 (2009).
[9] Radicchi, F. & Castellano, C. Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. *Journal of Informetrics* **6**, 121–130 (2012).

[10] Lehmann, S., Jackson, A. & Lautrup, B. Measures for measures. *Nature* **444**, 1003–1004 (2006).

[11] Bornmann, L. & Daniel, H. 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**, 427–440 (2006).

[12] Colwell, R. *et al.* Informing research choices: Indicators and judgment. Tech. Rep., The Expert Panel on Science Performance and Research Funding, Council of Canadian Academies (2012). URL http://www.scienceadvice.ca/uploads/eng/assessments%20and%20publications%20and%20releases/science%20performance/scienceperformance_fullreport_en_web.pdf.

[13] Radicchi, F., Fortunato, S. & Castellano, C. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences* **105**, 17268 (2008).

[14] Hirsch, J. An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences* **102**, 16569 (2005).

[15] Lundberg, J. Lifting the crown—citation z-score. *Journal of informetrics* **1**, 145–154 (2007).

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

[17] Hoang, D. T., Kaur, J. & Menczer, F. Crowdsourcing scholarly data. In *Proc. Web Science Conference: Extending the Frontiers of Society On-Line (WebSci)* (2010). URL http://journal.webscience.org/321/.

[18] Kaur, J. *et al.* Scholarometer: A social framework for analyzing impact across disciplines. *PLoS ONE* **7**, e43235 (2012). URL http://dx.doi.org/10.1371%2Fjournal.pone.0043235.
[19] Radicchi, F. & Castellano, C. A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE* 7, e33833 (2012). URL http://dx.doi.org/10.1371%2Fjournal.pone.0033833

[20] Mahmoud, H. *Polya urn models*, vol. 76 (Chapman & Hall/CRC, 2008).

[21] Redner, S. On the meaning of the h-index. *Journal of Statistical Mechanics: Theory and Experiment* 2010, L03005 (2010).

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

[23] Egghe, L. Theory and practise of the g-index. *Scientometrics* 69, 131–152 (2006).

[24] i-10 index. http://googlescholar.blogspot.com/2011/11/google-scholar-citations-open-to-all.html. URL http://googlescholar.blogspot.com/2011/11/google-scholar-citations-open-to-all.html

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

[26] Moed, H., De Bruin, R. & Van Leeuwen, T. New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics* 33, 381–422 (1995).

[27] Leydesdorff, L. & Bornmann, L. Integrated impact indicators compared with impact factors: An alternative research design with policy implications. *Journal of the American Society for Information Science and Technology* 62, 2133–2146 (2011).
Table A2: Top user-defined disciplines.

| Discipline                     | Authors | ⟨h⟩ |
|-------------------------------|---------|-----|
| 1. computer science           | 656     | 16.02 |
| 2. physics                    | 200     | 18.66 |
| 3. computer networks          | 130     | 16.25 |
| 4. bioinformatics              | 125     | 16.50 |
| 5. engineering                 | 115     | 11.46 |
| 6. medicine                   | 104     | 23.47 |
| 7. chemistry                  | 103     | 13.92 |
| 8. human computer interaction | 94      | 17.72 |
| 9. computer science, security | 82      | 19.32 |
| 10. image processing           | 80      | 18.39 |
Table A3: Manually clustered disciplines.

| Manual label                  | Disciplines                                                                 | Authors | h   |
|-------------------------------|-----------------------------------------------------------------------------|---------|-----|
| 1. computer science           | computer science, artificial intelligence image processing computer networks computer science computer science, theory & methods computer science, software engineering computer science, information systems computer science, hardware & architecture computer science, cybernetics | 4,342   | 15.79 |
| 2. biology                    | plant sciences biology zoology plant sciences evolutionary biology entomology biology biodiversity conservation biochemistry & molecular biology | 2,385   | 19.56 |
| 3. behavioral sciences        | sociology psychology, social psychology, applied anthropology psychology behavioral sciences | 1,846   | 17.97 |
| 4. engineering                | engineering, mechanical engineering, electrical & electronic engineering, biomedical | 1,302   | 14.93 |
| 5. economics                  | economics statistics & probability mathematics, applied mathematics | 972     | 17.02 |
| 6. mathematics                | political science public administration political science | 860     | 15.53 |
| 7. physics                    | physics, applied physics, multidisciplinary physics, condensed matter physics | 675     | 19.63 |
| 8. business                   | business, marketing management business, finance business | 665     | 15.59 |
| 9. education & educational research | education technology education & educational research | 305     | 12.18 |
| 10. humanities, multidisciplinary | humanities, multidisciplinary humanities | 122     | 9.00  |