Journal of Informetrics 8 (2014) 880–889

Contents lists available at ScienceDirect

Journal of Informetrics

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

The Matthew effect in economics reconsidered

Daniel Birkmaier[a], Klaus Wohlrabe[b,*]

[a] University of Munich, Germany
[b] Ifo Institute for Economic Research at the University of Munich, Poschingerstr. 5, 81679 Munich, Germany

ARTICLE INFO

Article history:
Received 17 May 2013
Received in revised form 29 July 2014
Accepted 18 August 2014
Available online 18 September 2014

JEL classification:
A12
A14

Keywords:
Matthew effect
Gibrat's law
Kolmogorov–Smirnov

ABSTRACT

We apply the test of Ijiri and Simon (1974) to a large data set of authors in economics. This test has been used by Tol (2009, 2013a) to identify a (within-author) Matthew effect for authors based on citations. We show that the test is quite sensitive to its underlying assumptions and identifies too often a potential Matthew effect. We propose an alternative test based on the pure form of Gibrat's law. It states that stochastic proportionate citation growth, i.e. independent of its size, leads to a lognormal distribution. By using a one-sided Kolmogorov–Smirnov test we test for deviations from the lognormal distribution which we interpret as an indication of the Matthew effect. Using our large data set we also explore potential empirical characteristics of economists with a Matthew effect.

1. Introduction

The *Matthew Effect* (Merton (1968)) has recently gained more and more attention in bibliometrics.[1] It states that papers are not only cited due to their quality but due to the fame of authors. Tol (2009) suggested using an approach by Ijiri and Simon (1974) to test this effect empirically. He showed that for 99 out of 100 economists, a within-paper Matthew effect was present. Tol (2013a) applied the test to different cohorts of economists using citation data from the RePEc website. Again, he found that the effect is statistically significant for all cohorts, being larger for older economists. In spirit of Tol (2009) we apply the test to a much larger data set of more than 10,000 economists using citations from RePEc. It turns out that the test is quite sensitive and detects a potential Matthew effect in more than 90% of the cases. Although the idea behind the test by Tol is intuitive and convincing, this might be an unrealistic scenario. Based on this result we discuss the robustness of the test and its underlying assumptions. The pure form of Gibrat's law, on which the test relies, leads to a lognormal distribution instead of a Pareto distribution as stated in Tol (2009). As an alternative test we propose to employ a Kolmogorov–Smirnov test to check whether the individual citation pattern of authors follows a lognormal distribution. We interpret deviations from the lognormal distribution as an indication of the Matthew effect. Our test detects considerably less authors with a Matthew effect. Furthermore, we show that such authors are not concentrated at the top, i.e. authors with the most citations. The paper is organised as follows: we first describe our data set which uses citations from RePEc. Then, we apply the test by Tol (2009) to our large data set and discuss its underlying assumptions and empirical regularities. After that, we outline some theoretical considerations and propose our new test which is followed by an exposition of the results of the alternative test. Finally, we investigate in a regression analysis if we can explain why some authors have a Matthew effect and some do not.

* Corresponding author. Tel.: +49 089 9224 1229.
E-mail address: wohlrabe@ifo.de (K. Wohlrabe).

[1] For a detailed review of the literature see Tol (2009) and Wang (2014).

http://dx.doi.org/10.1016/j.joi.2014.08.005
1751-1577/

Table 1
Descriptive statistics.

	Full Sample (n =10,564)					KS Sample (n =633)				
	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
Number of Works	50	36	42	11	1063	168	152	72	91	1063
Papers/ Year	3.2	2.7	2.0	0.4	35.5	6.8	6.1	2.9	2.5	29.5
Cited Works	32	23	28	11	494	118	105	43	81	494
Fraction non-cited	0.33	0.32	0.15	0.00	0.91	0.27	0.26	0.12	0.05	0.67
Citations	517	198	1162	21	32143	2952	1828	3163	195	32143
Citations/Year	26.1	14.1	42.5	1.5	1148.0	110.0	76.7	105.5	5.7	1148.0
Citations/Work	9.1	5.5	11.9	0.1	230.4	18.9	12.4	18.6	0.8	147.1
Self-Citation Rate	0.11	0.07	0.12	0.00	1.00	0.10	0.07	0.11	0.00	0.89
Years in the Profession	16	15	7	2	45	26	25	7	8	45
h-Index	9	7	6	2	83	24	22	11	6	83
Gini coefficient	0.56	0.56	0.11	0.14	0.94	0.64	0.64	0.08	0.36	0.85

This table reports descriptive statistics for various bibliometric indicators. *Number of Works* includes working papers, journal articles, chapters in books and monographs. The *Full Sample* refers to the sample employed in the next section. *KS Sample* represents the data used for the Kolmogorov–Smirnov test.

2. Data

As Tol (2013a), we use citation data from RePEc. In economics, RePEc (Research Papers in Economics, www.repec.org) has become an essential source for the spread of knowledge and ranking of individual authors and academic institutions. RePEc is based on the 'active participation principle', i.e. that authors, institutions and publishers have to register and to provide information to the network. This approach has the main advantage that a clear assignment of works and citations to authors and articles is possible.[2] Indeed, the RePEc story has become a success, with more than 36,000 registered authors with listed works and 12,000 institutions in economic sciences worldwide as of August 2013. RePEc has three main sources for extracting citations: First, it reads out all publicly available documents within the network. Due to missing (open) access to the article or technical problems it is not always possible to extract all citations. Second, archive maintainers may provide meta-information on citations for their journals. Third, registered authors can upload citations for every article. See Seiler and Wohlrabe (2012) for further details on RePEc.

Data were collected on 11 July 2013 using Citec.[3] We adapted and extended the Matlab code provided by Tol (2013b). We read out the publicly available citation profiles. These profiles contain, besides the aggregated citations, also citation numbers for each listed bibliographic item (books, working papers and journal articles). Furthermore we have the number of aggregated self-citations.[4] Based on the citations we calculate the h-index.[5] We assessed 35,633 citation pages.[6] For the upcoming analysis we only include authors who have more than ten papers with nonzero citations.[7] Furthermore, the sum of all citations should be at least 20.[8] This leaves us with 10,564 authors. In Table 1 we report some descriptive statistics for our sample of authors. We report all indicators which we were able to extract automatically from each individual citation page. The average number of works is about 20 and ranges between 11 and 1063. "Number of works" includes journal articles as well as books, chapters and working papers. As it is well known from the bibliometric literature, the distribution of citations is highly skewed across authors as the mean (517) and the median (197) differ substantially. "Years in the profession" denotes the time from the first publication until now. It shows that in our sample, at least two years have proceeded since the first publication and that the oldest author has published his first article 45 years ago. The Gini coefficient measures the citation concentration across papers for each author. If the coefficient equals 0 then each cited paper has the same citation count. In case of a Gini coefficient of almost 1, all citations are concentrated on one paper.

3. The test by Tol (2009) reconsidered

3.1. Empirical results for authors

The test of Tol (2009) for author i (within-paper) is given by

$$\ln Citations_i = \alpha_i + \beta_i \ln Rank_i + \gamma \ln^2 Rank_i. \tag{1}$$

[2] For instance, Google Scholar as a source for citation analysis potentially suffers from the problem of clear identification of citations which can lead to overestimation of citations, see Harzing and van der Wal (2009).
[3] See http://citec.repec.org/index.html for further details.
[4] Unfortunately, we were not able to extract the self-citations for each individual item.
[5] In Citec, citations are only consolidated across paper versions, not across document types. This may result in a bigger h-index than in the RePEc rankings.
[6] The full data set, including the individual citation profiles and summary statistics for each author, are available upon request from the authors.
[7] We assume that a within-paper Matthew effect for authors with only a few cited papers does not make sense. However, this does not rule out the possibility of a Matthew effect for one or two single papers with many citations.
[8] We want to exclude cases where 10 papers have one citation each.