

The scholarly communication of economic knowledge: a citation analysis of Google Scholar

Yutao Sun¹ · Belle Selene Xia²

Received: 6 June 2016 / Published online: 3 October 2016
© Akadémiai Kiadó, Budapest, Hungary 2016

Abstract Citation counts can be used as a proxy to study the scholarly communication of knowledge and the impact of research in academia. Previous research has addressed several important factors of citation counts. In this study, we aim to investigate whether there exist quantitative patterns behind citations, and thus provide a detailed analysis of the factors behind successful research. The study involves conducting quantitative analyses on how various features, such as the author's quality, the journal's impact factor, and the publishing year, of a published scientific article affect the number of citations. We carried out full-text searches in Google Scholar to obtain our data set on citation counts. The data set is then set up into panels and used to conduct the proposed analyses by employing a negative binomial regression. Our results show that attributes such as the author's quality and the journal's impact factor do have important contributions to its citations. In addition, an article's citation count does not only depend on its own properties as mentioned above but also depends on the quality, as measured by the number of citations, of its cited articles. That is, the number of citations of a paper seems to be affected by the number of citations of articles that the particular paper cites. This study provides statistical characteristics of how different features of an article affect the number of citations. In addition, it provides statistical evidence that the number of citations of a scientific article depends on the number of citations of the articles it cites.

Keywords Bibliometrics · Citation analysis · Economics

✉ Yutao Sun
sunyt100@nenu.edu.cn
Belle Selene Xia
belle.xia@helsinki.fi

¹ School of Economics, Northeast Normal University, Jingyue Dajie 2555, Changchun 130117, Jilin, China

² Department of Economics, University of Helsinki, Arkadiankatu 7, P.O. Box 17, 00014 Helsinki, Finland

Introduction

Many researchers who aim to study the spread of knowledge have previously used citation counts as a proxy. Jaffe and Trajtenberg (1998), for example, have investigated the international spreading of ideas using patent citations. Previous literature has also addressed several important factors of citation counts such as the authors' rankings and the journal's rankings (van Aalst 2010). In this study, we aim to investigate the statistical patterns of citations. Here the number of citations can be evaluated at the article level (e.g., how often an article was cited in other articles), journal level (e.g., average citation count for the articles in a journal), or author level (e.g., h-index).

For the purpose of our research, we focus on the article-level number of citations. In particular, we would like to establish a statistical model to explain the number of citations of an article by various features (defined below) of the article. We focus on research in the field of economics. In addition, we will introduce the concept of “cross-level effect” and define it to be the effect that the citation count of an article influences the citations of its citing papers. The research results presented in this study will be of key interests to the academic scholars in economics.

In the next section, we will present some evidence behind citation analysis, establishing the approach of using citation counts to represent the spreading of academic ideas. We will also present the research questions of our study. Following the research method section, we will detail our results along with their implications for future research. Finally, we will conclude our study in the last section of this paper.

Traceable spreading of knowledge in citation analysis

The cross-level effect of citation counts

Citation analyses provide important insights about (scholarly) communications of scientific findings in academia (Ding et al. 1999; Bornmann and Hans-Dieter 2008). In addition, Dewey (2016) points out that citation analyses have shown growing significance in terms of the quality of the publication and the reputation of the institutions. Yet, little research has been carried out on the roles of citations, the quality of the author of the paper, as well as the impact factor. In the present study, we will be looking at the cross-level effect as well as the quality of the publication on the citation counts in Google Scholar.

Dong and Schafer (2011) point out that the use of internet has significantly facilitated the access of scientific information as well as multidisciplinary cooperation, changing the way information is accessed and processed. This also results in an increasing in the production of scholarly research. This trend has led to the demand to specialize knowledge and to seek diverse sources of information. Therefore, improved search functions seem to play an important role, helping researchers to retrieve the information they need (Barnett et al. 2011). Citation counts are admittedly to be one important criterion for scholar to find relevant and reasonable information.

Previous research has shown that scholars nowadays tend to cite distantly related scientific research more than before. Kang et al. (2015) argue that this is a result of the development of information technologies and the increasing focus on interdisciplinary research. Citing closely related research papers has the aim of utilizing expertise knowledge related to one's own research area, whereas citing more distantly related research

papers has the aim of seeking out useful knowledge that is distant from one's area of expertise. The first type of citation can be used to refine existing ideas and the latter aims to generate new innovated ideas. Mingers and Lipitakis (2010) find that the current trend is that scholars are shifting their focus to a diverse source of information that may or may not be closely linked to their own field. In this paper, we will show that the citation of a particular publication does not only depend on its own authors, journal, and publication time but also depends on the articles it has cited. We denote this as the cross-level effect.

Focuses have been previously put on the determinants behind the citation counts in different journals. Examples of determinants include the year-wise distribution of citations, the rate of citations per article, the different type of documents cited, and the authorship pattern of the citations (Bauer and Backkalbasi 2005). As a response to the availability of a large set of knowledge, the goal of citation analysis is to recover useful patterns from the databases. In other words, the accurate extraction of citations is increasingly important. Kousha and Thelwall (2007) have provided discussions on several citation extraction and processing techniques. Davis and Fromerth (2006) argue that citations reveal important information about the publication and hence, may be used to distinguish the publications from each other.

On the other hand, researchers such as Leydesdorff (1998) and MacRoberts and MacRoberts (1989) have classified the distribution of knowledge to two types: the untraceable and traceable spreading of knowledge. The first type is completed primarily by means such as “obliterated by incorporation” (OBI, see also Garfield 1975). As an example, we expect currently to find scarce citations of the original source of “one plus one equals two”. OBI by name suggests that this type of communication cannot be included in quantitative studies. The second type, on the other hand, can be achieved by formal citations which are traceable in the bibliography of an article. This suggests that the citation is somewhat different from the communication of knowledge as a whole, since that only the traceable part can be included in studies. In addition, analyzing the traceable part is itself restricted: Citation analysis seems to be affected by a certain bias (MacRoberts and MacRoberts 1989). The process of citation has built a large network, of which the nodes are articles and the communication operators are the citations (Leydesdorff 1998). This network includes both listed and unlisted articles. Moreover, it might not be uniformly distributed in the sense that different articles attract a different number of readers. Even within a database, the distribution varies from fields to fields (Harzing and Wal 2008). This type of heterogeneity is intuitive: Ideas tend to move much faster in an active and practical field than an inactive and theoretical field.

The unique traceability together with the existence of various kinds of academic databases, which keep records on citations of a huge amount of articles, indicates that the citation counts is the only possible indicator for the communication of knowledge. One of the cautions in citation analysis is that, due to the uneven distribution of the citation network, citation analysis should be only applied to a particular field, aiming at discovering patterns of certain databases instead of predicting the properties of the entire network. This explains our choice of focusing on the field of economics. It is acknowledged that citation practices vary between natural science and social science fields and even between different areas within a single subfield.

The quality of publication

Vaughan and Shaw (2003) point out that the bibliometrics was originally developed to trace scientific ideas, the progress of science, and the impact of the work. Therefore,

bibliometrics also plays an important role in order to study how research ideas and discoveries are communicated internationally. Originally, citation counts were invented to understand the impact of scientific discoveries and research function. That is, citation counts were used to understand how scientific results in different fields are communicated (Harter and Ford 2000). They were not meant to rank any individual scientists, universities, or any other academic institutions.

Nowadays, citation counts have been used as a measure of quality as well as a measure of the productivity of the researchers so that the top cited researchers are recognized as being more productive and influential in their work than the less cited ones. Consequently, researchers in certain countries are obliged to have publications in certain journals. This will naturally benefit universities that use English as the primary language (Kousha and Thelwall 2008). It is also natural that citation counts show a heterogeneity in the sense that the number of citation for an article is different in economics than in, e.g., biology (Kayvan 2009). Moreover, the citation counts for books and journal articles differ when benchmarking authors and countries. This indicates that citation analysis for journal articles is not always informative for other types of publications (Kayvan and Abdoli 2010). In addition, the citation counts are used as an important measure to rank the top scientists and to discover the most widely cited ideas and scientific findings. However, the link between the quality of a publication and the citation counts is not very clear (Norris et al. 2008).

When finding the factors that contribute to the number of citations, Leydesdorff (1998) suggested that citations can be viewed as an indicator of reward in the academic system. If this indicator is to be decomposed, one possible element would be the “impact factor”. The impact factor is originally developed as a way of describing the importance of a journal, and is defined as the ratio of the number of citations in the current year to the number of articles published in the previous 2 years (Garfield 2006). In fact, the impact factor can be applied to authors, journals, and fields of study. The use of the impact factor has been also criticized. Researchers, such as Bordons et al. (2002), argue that impact factor depends heavily on the fields: Journals in popular fields are likely to have higher impact factor than others, while this does not necessarily imply smaller contributions of those having lower impact factor. Nevertheless, Amin and Mabe (2000) argue that impact factor is a prominent indicator for the significance of journals. Other instructive factors exist, such as author ranking and publishing year (van Aalst 2010).

Citation counts in Google Scholar

Google Scholar is a well-known database for studying the impact of citation sources and scholarly communication. The database provides useful information on the different characteristics of an article including, the document type, the language, the file format, the internet location, the number of citations, authors, and the publication year. In order to study different factors behind citation counts, previous research has examined the following research questions:

- What is the document type distribution of Google Scholar citing sources?
- What language, publication year and accessibility level (open access/non-open access) do Google Scholar citations have?
- What are the file format and Internet domain of the citing resources in these databases?
- How long does it take a published article to be cited by another article in the scholar database?
- How does the citation appear in the reference section of the citing resources?

In the present study, we select Google Scholar to be our primary source databases. Google Scholar has some unique properties. It is composed of an almost infinitely large collection of all databases available on the internet (Smith 2006). With this collection, Google Scholar is able to create an inter-database citation relation map for a particular article. The consequence of these inter-database citation links is that, for the same article, Google Scholar generally gives more citation counts than other databases such as the Web of Science (Harzing and Wal 2008). However, Google Scholar does truncate data in a way that only the first 1000 results matching the criteria will be returned, although the citation frequencies are calculated based on the entire collection (Smith 2006). Also, Google Scholar uses a complicated way of ordering the searching result: the results are ordered by using more than 200 signals including PageRank. In practice, when the search results are returned, they will be ordered by both “relevance” and “importance”. “Relevance” can be understood as the match with the searched keywords. The key of “importance” is the PageRank algorithm developed by Google Scholar itself. The PageRank measures the importance of a web page by the numbers of links, from other important web pages, towards it (van Aalst 2010). In addition, Google Scholar records articles that do not exist online. These may be working papers, unpublished articles or articles in private databases. Lastly, an article published later might be marked as a cited paper of an article published earlier. Articles showing this pattern are mostly books, especially textbooks.

Our study focuses on different aspects of Google Scholar. In particular, we have formulated the research questions of our study as follows:

1. Are the citations of an article statistically related to the citations of its cited papers, i.e., does the cross-level effect exist?
2. Are the number of citations of an article statistically related to the quality of its author, journal and publishing year?
3. If there is indeed a statistical relation between these attributes and the number of citations, then how exact do they relate, i.e.: Is there an order in the importance of these attributes?

Research methods

Data collection

We obtained two samples, in the year of 2011, on the citations of journal articles from Google Scholar by searching with the keyword “economics OR econometrics OR finance”. The first sample is a collection of 21,807 articles published in various journals. Here in this sample, we were not able to identify the name of every journal on which the articles are published. The second sample is a set of 4919 articles collected from five top-ranking journals including the Journal of Economic Literature, the Journal of Financial Economics, the Quarterly Journal of Economics, the Journal of Political Economy, and *Econometrica*. We selected these five journals as they are amongst the top journals in the fields, and hence would help us analyze the citations of articles that are published in top journals. These five journals are merely samples of top journals and there are, of course, many other top journals in the fields. However, as we already have a sufficient amount of data from these five journals and as top journals are similar in some sense (e.g., rigorous peer-review, high impact factors, etc.), we argue that the inclusion of other top journals would not contribute significantly to our study in a statistical perspective. In addition, the

data collection process would become more time-consuming if more journals are to be included. If one would like to investigate the same objective on other journals, one could simply repeat the same approach that we are going to use on those journals when data becomes available.

The data collection procedure of the first sample is the following. We first collect 100 articles from the search result. We call these the level-0 articles. Next, for each of these level-0 articles, we collect another 100 papers, which cite the article. That is, for an article A, we collect 100 papers that cite article A, and this procedure is repeated for all of the 100 articles collected in the first step. In this step, the total number of articles collected is, naturally, 10,000. We call these the level-1 articles. Now for each of these 10,000 level-1 articles, we collect another 30 papers that cite the article. In this step, we were able to collect 300,000 papers in total. We call these the level-2 articles. The sample containing these 300,000 articles is then cleaned so that an article is dropped if the publishing journal or publishing year is missing, or if it is marked as “[Book]” by Google Scholar. In addition, we remove the journals with less than 50 articles in our sample. This is to improve the quality of the estimation and the tests. This procedure gave 21,807 papers in total that formulate the first sample, the recursive sample. Here the set of all level-0 to level-2 articles is organized in a way such that a citation tree can be built. This tree has a clear structure: the articles in level-0 are called the roots, whereas the articles in the last level are called the ends. The remaining articles (the level-1 articles) are called the nodes. Citation relationships, or simply parent–child pairs, link all the points. The roots can only be parents and the ends can only be children whereas the remaining articles can be parents or children. In this paper, we only use the level-2 articles for the analysis, taking the numbers of citations of the corresponding level-0 and level-1 articles as independent variables (covariates). For the second sample, the flat sample, the collection procedure is simpler. For each of the five journals listed previously, we collect 1000 articles from the search result. These 5000 articles are then cleaned in a similar way as the first sample, resulting in 4919 articles, which formulate the flat sample.

After the initial collection, we set up these samples into panels where the journals are the groups. For each observation, we have the following variables.

- Number of citations: the number of citations of the article.
- Impact factor: The impact factor (in the year of 2011) of the journal on which the article is published.
- Number of versions: The number of versions of the article as reported by Google Scholar. The same article being in different databases leads to Number of Versions being larger than 1.
- Age of paper: The age of the article as standing in the year of 2011. The value is calculated as 2011 minus the actual publishing year of the article.
- Author score: The score of the first author. The value is calculated as the logarithm of the difference between 1,391 and the first author’s ranking obtained from the “Economist Rankings at IDEAS” in the year of 2011, where 1391 is the number of authors for whom we were able to obtain the ranking data. Here the higher the score is, the better the author is according to the ranking. When the author is not found in our ranking data, Author Score will be set to 0.
- Parent citations 1: The number of citations, in logarithm, of the corresponding level-1 article.
- Parent citations 2: The number of citations, in logarithm, of the corresponding level-0 article.

- Is citation: A binary variable taking value 1 indicating the full text of the article cannot be found in any database, and hence is recorded as “[Citation]” by Google Scholar.
- Author missing: A binary variable taking value 1 if the author is missing on the search result or if the author is not registered in the “Economist Rankings at IDEAS”.

Here Parent Citations 1, Parent Citations 2, and Is Citation are only for the recursive sample. Table 1 presents a summary of the two samples.

A quick look at the descriptive statistics gives the following findings. First, the mean number of citations in the flat sample is much higher than that of the recursive sample. This is natural since the flat sample is constructed only on top journals, and the articles in these journals are usually cited more. Second. The mean of the author score is also higher in the flat sample. This is also intuitive as top journals have more publications from good authors. Third, the mean of the impact factor is also larger. This is natural since journals are considered top when their impact factor is large.

In Table 2, we present the number of articles corrected from each of the five journals in the flat sample.

Modeling methods

We would like to infer how the attributes of an article affect the number of citations. That is, to model Number of Citations, the dependent variable, on the rest of the variables, the regressors, described above. We choose the fixed-effect negative binomial model for our inference.

Given the nature of Number of Citations, the econometric tools modeling discrete counts can be invoked. These tools include the ordinary Poisson regression (see Cameron and Trivedi 2013, chap. 3), the Poisson regression with quasi-likelihood function (the quasi-Poisson, see Pesaran and Malden 1997, chap. 8), the negative binomial regression (see Hilbe 2011), and the zero-inflated Poisson/negative binomial regression (see Lambert 1992; Greene 1994). For the estimation procedures, as we work with panels, the pooled estimation, the fixed-effect estimation, and the random-effect estimation are all candidate procedures for the inference.

As a comparison of models, the Poisson regression relies on the strict assumption of equal conditional mean and conditional variance of the dependent variable. When this assumption is violated, the parameter estimates will be consistent but the standard errors of these estimates will be incorrect (see Ismail and Jemain 2007). The quasi-Poisson and the negative binomial regression allow the relaxation of this assumption, and hence, allow the presence of an overdispersion in the dependent variable. Here quasi-Poisson will deliver the same parameter estimates as the ordinary Poisson and the standard errors of these estimates will be correct also (hence will be different from Poisson). The negative binomial regression accommodates the overdispersion by estimating an additional parameter called the overdispersion parameter. The zero-inflated Poisson regression is modified on top of the ordinary Poisson regression to allow for a situation called zero inflation. This is useful when there is an excessive presence of zeros in the dependent variable. However, the zero-inflated Poisson itself does not allow for the presence of the overdispersion. The zero-inflated negative binomial, however, allows for the presence of both the overdispersion and the zero inflation. For our samples, the standard deviation of Number of Citations is much greater than the mean, indicating a sign of overdispersion in Number of Citations. This may hint the usage of the ordinary Poisson model inappropriate. On the other hand, the number of articles with zero citations is small relative to the total number of observations.

Table 1 Descriptive statistics

Variables	Recursive sample		Flat sample	
	Mean	SD	Mean	SD
Number of citations	71.3145	227.7053	299.0307	738.1900
Impact factor	1.7242	1.3394	4.8919	1.1255
Number of versions	12.9186	9.7409	12.0992	10.4363
Age of paper	5.2852	4.7466	18.5924	12.9040
Author score	0.7338	2.0745	1.7015	2.9292
Parent Citation 1	5.2065	1.1656	–	–
Parent Citation 2	7.4189	0.7809	–	–
Count of Zero citations	1284		463	
Number of observations	21,807		4919	

Parent Citation 1 and Parent Citation 2 are only available in the recursive sample

SD stands for the standard deviation

Table 2 Number of articles in flat sample

Journals	Number of articles
Journal of economic literature	999
Journal of financial economics	1000
Quarterly journal of economics	974
Journal of political economy	962
Econometrica	984
Total number of articles	4919

This indicates that a zero inflation is likely not the case so that the zero-inflated Poisson/negative binomial regression is not necessary. With these observations, we are left with the quasi-Poisson regression and the negative binomial regression. We will analysis the choice from the two later.

For the differences of the estimation procedures, the pooled estimation requires that the observations are homogeneous, whereas the fixed-effect and the random-effect estimation both allow for the unobserved heterogeneity in the observations. The difference is that the fixed-effect estimation allows for a correlation between the unobserved heterogeneity and the regressors. For our sample, it is likely that each journal may have an unobserved heterogeneity, or characteristic, that captures, e.g., the underlying quality of the journal. This unobserved characteristic is likely to be correlated with Impact Factor, as the impact factor is one of the measures of this characteristic (the quality of the journal). This renders the usage of a pooled or a random-effect estimation inappropriate in our setting. Here alternatively, it would be possible to drop Impact Factor and estimate consistently the models with random effects. In that case, however, we would not be able to capture the effect of Impact Factor on Number of Citations, which is one of our objectives. For these reasons, we will invoke the fixed-effect estimation as out main inference method. We will also report the estimation results from the pooled and the random-effect variants of these models for comparison and for Hausman (1978) specification tests.

In addition, when working with a fixed-effect estimation, we need to investigate the so-called incidental parameter problem (see Neyman and Scott 1948; Lancaster 2000). In the presence of the incidental parameter problem, certain parameter estimates can be asymptotically biased. Here the fixed-effect negative binomial regression is partially subjected to the incidental parameter problem, in the sense that the overdispersion parameter is estimated with an asymptotic bias (see Allison and Waterman 2002). This does not impose any problem since the overdispersion parameter is not of interest for what we concern. However, it would be inappropriate to explicitly model and explain the overdispersion using any regressors, as the estimates of the coefficients associated with these regressors will be asymptotically biased. Furthermore, the incidental parameter problem also affects the zero-inflated Poisson/negative binomial regression. These two models incorporate a component developed from the logit or the probit regression to model the zero inflation. However, it is known that the logit and the probit model themselves are subjected to the incidental parameter problem (see Chamberlain 1980; Fernández-Val 2009 respectively). This indicates that the zero inflation cannot be modeled correctly in a zero-inflated Poisson/negative binomial regression. This incorrectness will also affect the inference that we concern in this paper. For the logit model, Cox (1958) has developed a solution to logit model to eliminate the incidental parameter problem. However, it has not been studied if this solution is also applicable for the zero-inflated Poisson/negative binomial regression.

Next, we explain the choice between the quasi-Poisson and the negative binomial. Whereas they are both candidate models for our analysis, we exclude the quasi-Poisson because the Hausman test would not be applicable for this model, in the sense that the test-statistic will not be χ^2 -distributed even under the null hypothesis. Without the Hausman test, we would not be able to distinguish the fixed-effect quasi-Poisson against the random-effect quasi-Poisson.

As a summary, we choose the fixed-effect negative binomial model (without modeling the overdispersion by regressors) as our main model. We will also report the estimation results from the pooled and the random-effect negative binomial model for comparisons and for specification test.

Furthermore, for each of the two samples we have, we will carry out the estimations twice, on the full sample and on the identified subsample. The identified subsample, on the other hand, is a subsample formulated by including only articles (from the full sample) whose author is identified in the “Economist Rankings at IDEAS”. This allows a comparison with the full sample.

Estimation result

Tables 3 and 4 report the estimation results of the fixed-effect negative binomial regression on, respectively, the full sample and the identified subsample. Significant estimates are reported in bold face. In these tables, we also report the Hausman test of the fixed-effect model versus the pooled model, taking the fixed-effect model as the consistent model (FE vs. P) and of the fixed-effect model versus the random-effect model (FE vs. RE). Here note that, under the identified recursive sample, the Hausman test-statistic of the fixed-effect model versus the pooled model was initially negative in our study. However, we reference our work from Schreiber (2008) to just use the absolute value of the test-statistic to construct the test.

A first look gives the following finding. The fixed-effect model indeed is preferred against the pooled and the random-effect model according to the tests. In fact, the tests

Table 3 Estimation result from negative binomial regression—full sample

Variables	Recursive sample			Flat sample		
	FE	RE	P	FE	RE	P
Impact factor	0.116 (0.057)	-0.065 (0.007)	0.242 (0.006)	0.081 (0.034)	-0.335 (0.015)	-0.200 (0.017)
Number of versions	0.062 (0.001)	0.026 (0.001)	0.064 (0.001)	0.082 (0.003)	0.032 (0.001)	0.076 (0.003)
Age of paper	0.177 (0.002)	0.057 (0.001)	0.169 (0.002)	0.056 (0.002)	0.020 (0.001)	0.054 (0.002)
Author score	0.093 (0.028)	0.158 (0.020)	0.105 (0.029)	0.101 (0.058)	0.196 (0.041)	0.126 (0.060)
Parent citation 1	0.253 (0.007)	0.337 (0.006)	0.276 (0.007)	–	–	–
Parent citation 2	0.055 (0.010)	0.030 (0.008)	0.073 (0.010)	–	–	–
Likelihood-ratio test	16,383.48	21,598.38	20,913.10	1723.22	3747.78	1843.65
Degree of freedom	8			5		
FE versus P	5688.07			176.90		
FE versus RE	9500.07			643.72		
Number of journals	119			5		
Mean #Articles per journal	183			983		
Obs	21,807			4191		

Bold face indicates significance at the 5 % level. Each element reports the estimate of the coefficient associated with the variable in the first column. The standard error is in parentheses. FE, RE, and P represent the fixed-effect, the random-effect, and the pooled estimation respectively. Other control variables: Author missing and Is Citation (for the recursive sample only). FE versus P reports the Hausman test-statistic of the fixed-effect model versus the pooled model, taking the fixed-effect model as the consistent model. FE versus RE reports the Hausman test-statistic of the fixed-effect model versus the random-effect model, taking the fixed-effect model as the consistent model

reject the equivalency of the fixed-effect model to the pooled model and of the fixed-effect model to the random-effect model with strong evidence, indicating that, in our context, the fixed-model is the only consistent model. This validates our model selection.

In addition, this statistically demonstrates the general opinion that the impact factor is a measure of the underlying (unobserved) characteristic of a journal, but is not an error-free measure. If the impact factor did not reflect (i.e., were uncorrelated with) the underlying characteristic of a journal at all, then the fixed-effect and the random-effect model would be very similar on the estimates of Impact Factor. However, if the impact factor were an error-free measure, in which case there would be no underlying journal-specific characteristics, then the fixed-effect model and the pooled model would be very similar.

Furthermore, we also statistically demonstrated that citation is a channel for the transfer of quality of publications, i.e., the existence of the cross-level effect. Recall that we argue the number of citations is a measure of the quality of a publication. Here the coefficient estimates of Parent Citation 1 and Parent Citation 2 are significant, indicating that the quality (i.e., the number of citations) of an article is positively related to the quality (the number of citations) of the article it cites and is also positively related to the quality of the

Table 4 Estimation result from negative binomial regression—identified subsample

Variables	Identified recursive sample			Identified flat sample		
	FE	RE	P	FE	RE	P
Impact factor	0.350 (0.005)	-0.085 (0.016)	0.191 (0.013)	-0.070 (0.054)	-0.217 (0.028)	-0.108 (0.230)
Number of versions	0.052 (0.002)	0.029 (0.001)	0.058 (0.002)	0.066 (0.005)	0.031 (0.001)	0.068 (0.005)
Age of paper	0.155 (0.005)	0.052 (0.002)	0.143 (0.005)	0.045 (0.004)	0.020 (0.002)	0.050 (0.004)
Author score	0.114 (0.029)	0.146 (0.023)	0.135 (0.029)	0.142 (0.053)	0.165 (0.043)	0.156 (0.053)
Parent Citation 1	0.231 (0.020)	0.278 (0.016)	0.246 (0.020)	–	–	–
Parent Citation 2	0.004 (0.030)	0.058 (0.023)	0.019 (0.030)	–	–	–
Likelihood-ratio test	1866.40	2968.56	2489.77	389.60	1233.68	477.61
Degree of freedom	7			4		
FE versus P	38.86			16.00		
FE versus RE	592.23			74.95		
Number of journals	117			5		
Mean #Articles per journal	21			252		
Obs	2455			1261		

Bold face indicates significance at the 5 % level. Each element reports the estimate of the coefficient associated with the variable in the first column. The standard error is in parentheses. FE, RE, and P represent the fixed-effect, the random-effect, and the pooled estimation respectively. Other control variables: Is Citation (for the recursive sample only). FE versus P reports the Hausman test-statistic of the fixed-effect model versus the pooled model, taking the fixed-effect model as the consistent model. FE versus RE reports the Hausman test-statistic of the fixed-effect model versus the random-effect model, taking the fixed-effect model as the consistent model

(level-2) article that the cited (level-1) article itself cites. This transfer, however, is not lossless, as the coefficient estimate of Parent Citation 2 is smaller than that of Parent Citation 1. In the identified recursive sample, however, the coefficient estimate of Parent Citation 2 is insignificant. This could be that an identified (hence “good” in some sense) author may be more independent in the sense that the quality of the author himself contributes to the number of citations of his article more than the cited articles do.

On the other hand, Impact Factor, Number of Versions, Age of Paper, Author Score all contributes to Number of Citations positively. This is in line with our expectation: Articles published in good journals, available in many databases, published earlier, or written by good authors are likely to receive more citations. However, it can be seen that the coefficient estimate of Impact Factor under the flat samples is smaller than that under the recursive sample. This is systematic across Tables 3 and 4. The reason could be that top journals already have a larger impact factor so that an addition increase in the impact factor would not lead to a large increase in the number of citations. In contraction with this, we see that the coefficient estimate of Number of Version is higher under the flat sample. This

is somewhat natural: Good journals are more likely to be included in databases that have more users. Therefore, articles published in these journals have more readers.

For recursive samples, when switching from the full sample to the identified subsample, a notable pattern can be observed: The coefficient estimate of Impact Factor increases dramatically whereas other estimates remain roughly the same. For this, we have no formal explanation. However, this may simply be a consequence that the identified subsample provides more information. On the other hand, for the flat samples when switching from full to identified, the coefficient estimate of Impact Factor becomes insignificant whereas the coefficient estimate of Author Score becomes significant. These together may indicate that the quality of the author contributes to the quality of the paper more than the impact factor of the journal does. This is realistic, as many good authors in the field we studied (such as W. Newey, M. Arellano, J. Hahn etc.) have unpublished articles that are frequently cited.

As a summary of the estimation results, we statistically demonstrated the general opinion that the impact factor is a measure of the underlying characteristic (quality) of a journal, but it is not an error-free measure. In addition, we make the following conjecture based on the estimation results. First, the quality of articles does transfer to other articles through citation. Second, the quality of an author plays a more important role to the quality of the paper this author writes, than the quality of the article cited. Third, the quality of an author is more contributive to the quality of his article, than the impact factor of the journal on which the article is published.

Conclusion and discussion

In this paper, we have taken the number of citations as a measure of the spread of knowledge in the field of economics and as a proxy to measure the quality of a published journal articles. We attempted to explain, statistically, how various factors affect the number of citations, and thus the spread of knowledge and the quality of an article. Specifically, we have randomly collected a data set of a number of published articles from Google Scholar, the well-known search engine for scientific articles in the area of economics and closely related. The data are collected in two ways: (1) without an in-advance constraint on the journals on which the articles are published or on the author of the article, and (2) with a prior constraint that the articles are published in (five) top journals in the field or are written by established authors. Subsequently, statistical estimations and analyses based on panel-data approaches are carried out to examine how the number of versions (as in different databases), the impact factor of journals, the age of the paper, and a score for the author that is constructed based on a publicly available ranking system.

The conclusions of our analyses are detailed as follows. First, the impact factor, the number of versions, the age of paper, and the author score all seem to contribute to the number of citations positively. Secondly, the impact factor is a measure of the underlying characteristic (quality) of a journal, but it is not an error-free measure. In addition, we conjecture that, based on our analysis, (1) the quality of articles does transfer to other articles through citation; (2) the quality of an author plays a more important role on the quality of the paper he or she writes rather than on the quality of the article cited; (3) the quality of an author is more contributive to the quality of his or her article rather than on the impact factor of the journal on which the article is published.

We would like to acknowledge the limitations of our studies. First, we argue that the impact factor is not an error-free measure of the quality of a journal. In general, this argument may also apply to, for instance, the author score we calculated, in which case a panel may also be set up for authors. However, from an empirical study point of view, this issue is admittedly difficult to solve statistically. As a matter of fact, it would be unrealistic to set up a panel according to different authors, as the number of authors is often far larger than the number of journals and the number of articles an author writes is perceived as smaller than the number of articles available in a journal; i.e., the number of observations would not be sufficient to set up a panel that allows an accurate inference. In addition, due to the existence of “OBI” and other implicit forums for knowledge sharing, it is not perfect to use citations as the proxy for the spreading of economic knowledge. Nevertheless, it is scientifically acknowledged to be one of the most explicit and quantitative measures so far that could be included in econometric research.

When it comes to future studies, it is worth to investigate the same research objective in other disciplines. In the field of economics, the number of authors of a single article is often limited, whereas, in other disciplines such as natural science, a single scientific article can have a large number of authors. In that case, the effect of the quality of a single author may, and most likely will, exhibit a completely different pattern to what we have presented in this paper. On the other hand, the same study may also be carried out to other search engines such as the Web of Science. Initially, we did study the Web of Science, in a non-panel setting. However, we found that the number of articles per journal was too small in our sample to set up a panel. It was due to this particular reason have we decided to exclude the related discussion.

Acknowledgments The authors would like to thank all the researchers and professors at KU Leuven who have shared their valuable comments with us. The research results presented in this study are based on the thesis work done at KU Leuven on citation analysis. The authors would also like to acknowledge the helpful and constructive comments presented by the anonymous reviewers, which has resulted in a major revision of this paper.

References

- Allison, P. D., & Waterman, R. P. (2002). Fixed-effects negative binomial regression models. *Sociological Methodology*, 32(1), 247–265.
- Amin, M., & Mabe, M. (2000). Impact factors: use and abuse. *Perspectives in Publishing*, 1, 1.
- Barnett, G., Huh, C., & Kim, Y. (2011). Citations among communication journals and other disciplines: a network analysis. *Scientometrics*, 88, 449.
- Bauer, K., & Backkalbasi, N. (2005). An examination of citation counts in a new scholarly communication environment. *D-Lib Magazine*, 11, 9.
- Bordons, M., Fernández, M. T., & Gómez, I. (2002). Advantages and limitations in the use of impact factor measures for the assessment of research performance in a peripheral country. *Scientometrics*, 53(2), 195–206.
- Bornmann, L., & Hans-Dieter, D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), 45–80.
- Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge: Cambridge University Press.
- Chamberlain, G. (1980). Analysis of covariance with qualitative data. *The Review of Economic Studies*, 47, 225–238.
- Cox, D. (1958). Two further applications of a model for binary regression. *Biometrika*, 45, 562–565.
- Davis, P. M., & Fromerth, M. J. (2006). Does the arxiv lead to higher citations and reduced publisher downloads for mathematics articles. *Scientometrics*, 71, 2.

- Dewey, S. H. (2016). (Non-) use of Foucault's archaeology of knowledge and order of things in LIS journal literature, 1990–2015. *Journal of Documentation*, 72(3), 454–489.
- Ding, C., Chi, C. H., Deng, J. & Dong, C. L. (1999). Citation retrieval in digital libraries, IEEE International Conference on Systems Man and Cybernetics. *IEEE SMC '99 Conference Proceedings*, Tokyo (vol. 2, pp. 105–109).
- Dong, C. & Schafer, U. (2011). Ensemble-style self-training on citation classification. *Proceedings of 5th International Joint Conference on Natural Language Processing*, (pp. 623–631). Chiang Mai, Thailand: Asian Federation of Natural Language Processing.
- Fernández-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics*, 150(1), 71–85.
- Garfield, E. (1975). The obliteration phenomenon. *Current Contents*, 51(52), 5–7.
- Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA: The Journal of the American Medical Association*, 15(305), 1511–1610.
- Greene, W. H. (1994). *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models*. New York: University Working Papers.
- Harter, S. P., & Ford, C. E. (2000). Web-based analysis of e-journal impact: approaches, problems, and issues. *Journal of the American Society for Information Science*, 51(13), 1159–1176.
- Harzing, A. K., & Wal, R. V. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, 8, 61.
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6), 1251–1271.
- Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge: Cambridge University Press.
- Ismail, N., & Jemain A. A. (2007). Handling overdispersion with negative binomial and generalized Poisson regression models. *Casualty Actuarial Society Forum*, 103–158.
- Jaffe, A., & Trajtenberg, M. (1998). *International knowledge flows: evidence from patent citations*. MA, USA: National Bureau of Economic Research Cambridge.
- Kang, M., Shin, J. D., & Kim, B. (2015). Automatic subject classification of Korean journals based on KSCD. *Indian Journal of Science and Technology*, 8(S1), 452–456.
- Kayvan, K. (2009). Characteristics of open access scholarly publishing: a multidisciplinary study. *Aslib Proceedings*, 61(4), 394–406.
- Kayvan, K., & Abdoli, M. (2010). The citation impact of open access agricultural research: a comparison between OA and non-OA publications. *Online Information Review*, 34(5), 772–785.
- Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: a multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, 58, 6.
- Kousha, K., & Thelwall, M. (2008). Sources of Google Scholar citations outside the science citation index: a comparison between four science disciplines. *Scientometrics*, 74(2), 273–294.
- Lambert, D. (1992). Zero-inated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1), 1–14.
- Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2), 391–413.
- Leydesdorff, L. (1998). Theories of citations. *Scientometrics*, 1(43), 5–25.
- MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: a critical review. *Journal of the American Society for Information Science*, 5(40), 342–349.
- Mingers, J., & Lipitakis, E. (2010). Counting the citations: a comparison of Web of Science and Google Scholar in the field of business and management. *Scientometrics*, 85(2), 613–625.
- Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1), 1–32.
- Norris, M., Oppenheim, C., & Rowland, F. (2008). The citation advantage of open-access articles. *Journal of the American Society for Information Science and Technology*, 59(12), 1963–1972.
- Pesaran, M., & Malden, P. S. (1997). *Handbook of applied econometrics* (Vol. 2). Oxford: Blackwell Publishers.
- Schreiber, S. (2008). The Hausman test statistic can be negative even asymptotically. *Jahrbücher für Nationalökonomie und Statistik*, 228(4), 394–405.
- Smith, A. G. (2006). Google Scholar as a cybermetric tool: a comparison with the New Zealand PBRF research assessment, *9th International Conference on Science and Technology Indicators*. Leuven, Belgium.
- van Aalst, J. (2010). Using Google Scholar to estimate the impact of journal articles in education. *Educational Researcher*, 39(5), 387–400.
- Vaughan, L., & Shaw, D. (2003). Bibliographic and Web citations: what is the difference. *Journal of the American Society for Information Science and Technology*, 54(14), 1313–1322.