# Online accessibility of scholarly literature, and academic innovation*

Timo Boppart
IIES, Stockholm University

Kevin E. Staub
University of Melbourne

May 10, 2016

**Abstract**

Starting in the late 1990s, preexisting volumes of economics journals were scanned and uploaded to the internet, making these articles accessible online via search engines and hyperlinks. This paper analyzes the effect of this online accessibility of the literature on the innovational strength of follow-on research. We provide a new measure of innovational strength that depends on the novelty of the combinations of cited references. This measure quantifies whether an article relies on references that are already well connected in the citation network or rather combines different strands of the literature for the first time. We document that online accessibility led to more innovative follow-on research according to this measure. The same qualitative effects are found with an alternative output measure of creativity based on the number of topics an article touches on. Moreover, we show that our measure of innovational strength has predictive power for received future citations.

**Keywords:** Digitization, online publication, bibliometrics, knowledge production function, recombinant growth, citations, networks, scholarly communication.

**JEL classification**: A11, D83, O31, O33.

* *Acknowledgments*: For helpful comments and suggestions we thank Hartmut Egger, Josef Falkinger, Mark McCabe, Ignacio Palacios-Huerta, Doh-Shin Jeon, Chris Snyder, Rainer Winkelmann, and seminar participants at the NBER Workshop on "Scholarly Communication, Open Science and Its Impact", the conference "The Economics of Information and Communications Technologies" in Paris, the Swiss Economist Abroad meeting in Lucerne, the TIGER Forum in Toulouse, the German Economic Association meeting in Goettingen, the European Economic Association meeting in Gothenburg, and the International Panel Data Conference in Budapest, as well as seminar participants at the universities of ANU, Bayreuth, Bern, Calgary, Carleton, Deakin, EIEF, Essex, Melbourne, Monash, New South Wales, Otago, Southern Denmark and Zurich. For providing us with data and for their explanations we are very thankful to Mary Glose from FSO, Peter Vlahakis and Luke Hospadaruk from JSTOR, Steve Husted from EconLit, and Mark McCabe. Nila Chea, Barbara Klotz and Lukas Rohrer provided excellent research assistance. A special thanks to Christian Elsasser for his computational help. Timo Boppart thanks the Knut och Alice Wallenbergs stiftelse KAW 2011.0144. for financial support. Kevin Staub gratefully acknowledges support by the Swiss National Science Foundation through fellowship PBZHP1-138692. An older version of this paper circulated as "Online accessibility to academic articles and the diversity of economics."
*E-mails for correspondence:* timo.boppart@iies.su.se, kevin.staub@unimelb.edu.au

# 1 Introduction

*"Among chosen combinations the most fertile will often be those formed of elements drawn from domains which are far apart." (Henri Poincaré, 1910, p. 325)*

How did the internet affect the academic process of knowledge accumulation? In this paper we focus on one particular aspect of the internet, namely the possibility to access preexisting journal articles online. This online accessibility allowed researchers to use powerful new tools (such as hyperlinks or search engines) which changed literature inquiries and, consequently, the process of knowledge accumulation. Analyzing publication data of 50 economics journals we assess empirically how online accessibility of the existing literature affected the innovational strength of follow-on papers.

Does online accessibility allow researchers to find some of the literature's lost pearls? Does it consequently foster innovation and creativity? Or do instead some inherent popularity traps lead to a Balkanization of the literature? To answer these questions, we propose a novel measure of innovational strength which relies on the geodesic (i.e., shortest) distance of cited references in the entire citation network. More specifically, using a network of citations, we calculate for each publication whether the references made cite each other directly, or via one or more additional documents. This measures whether a citing publication is innovative in the sense that it combines different strands of the literature for the first time, or whether instead it draws only from a relatively narrow, already well-connected literature. We exploit variation in the date at which different journal volumes have been made accessible online to estimate the effect of online accessibility on the innovational strength of follow-up articles. Based on the set of cited journals we calculate for each article the fraction of its 'relevant literature' that has been online previous to its publication. Since the relevant literature is article-specific, the constructed measure of online accessibility differs in the cross-section. Our identification strategy exploits this cross-sectional variation, allowing us to control for time trends in a flexible form. Our results show that online access leads authors to connect the literature in a more innovative way. The effect of online accessibility of the literature is not only reflected in terms of inputs (innovativeness or creativity of the combination of references). Quantitatively similar effects are found for the number of fields and subfields a paper touches on. Moreover, we document that our measure of innovational strength predicts the number of citations received in the long-run future, suggesting that online accessibility indeed improved the quality of follow-on research.

In this paper we analyze the data through the lens of a recombinant growth framework (Weitzman, 1996, 1998a). Scientific articles are considered as both input and output of the knowledge accumulation process. An article (or 'idea') is generated as a combination of preexisting articles.

The key determinant of an article's quality is the creativity of this newly formed combination. Citation data allows us to observe on which existing publications an article relies on, and we use information from the entire network of citations to determine the degree of innovation or creativity of an article's combination of existing publications. In recombinant growth theory, the 'diversity' of the pool of accessible articles is important since it determines the information content of the literature.[1] Thus, this framework emphasizes the preservation and storage of preexisting knowledge to facilitate innovative combinations.[2] In this paper we assess empirically whether platforms of online access improve this process.

As pioneered by Evans (2008), our paper analyzes the arguably exogenous variation in the date different journal volumes have been scanned and made accessible online. Evans (2008) documents that as more scientific articles became available online, more recent articles were referenced more often and citations were more concentrated on fewer documents. There is a small, recent literature which explores the impact of online access for the economics and business profession (Depken & Ward, 2009; McCabe & Snyder, 2015). So far, this literature has focused primarily on the impact an article's online accessibility had on its number of received citations. Depken & Ward (2009) show that access to the online platform Journal STORage (JSTOR) increases the number of citations to journals contained in JSTOR as well as to older journal volumes. McCabe & Snyder (2015) document that, after flexibly controlling for article quality, online accessibility had no overall effect on the number of citations a publication receives; an exception is the platform JSTOR where they find increases of about 10 percent. The authors also show that this increase is about the same for both often-cited as well as rarely-cited papers. Our paper addresses quite a different aspect of the scientific process by studying the impact of online accessibility of the academic literature on characteristics of follow-on research. In this respect, our research question is closer to Evans (2008). However, compared to the aggregate diversity measures considered by Evans (2008), our outcome variable measures the degree of innovation at the article level. Our identification strategy indeed exploits cross-sectional variation at the article level which allows us to disentangle the effect of online accessibility from any time trend.

Our paper is also related to Murray *et al.* (2009), who document how a reduction in intellectual property restrictions on a certain type of genetically engineered mice changed the level and type of (citing) follow-on studies. The reduction in intellectual property restrictions results in a more creative use of this type of genetically engineered mice, by new authors and institutions, in new

[1] Weitzman (1992, 1998a). See also Acemoglu (2011) for a theoretical characterization of the optimal level of diversity in research.

[2] Or as Weitzman (1998b, p. 333) puts it: "[T]he ultimate limits to growth may lie not so much in our ability to generate new ideas, so much as in our ability to process an abundance of potentially new seed ideas into usable form." See Ghiglino (2012) for a recombinant growth model which incorporates a mechanism to process such 'seed ideas'.

journals, and in the context of new keywords. In our paper, the exogenous shift in openness comes from the fact that preexisting papers have been scanned and made accessible online. As we document, this increase in openness also led to a more creative follow-on research.

The paper is structured as follows: In Section 2, we show how our new measure of an article's innovational strength is constructed and we highlight some empirical evidence of its predictive power of long-term citations. Section 3 describes our measure of online accessibility of the academic literature, as well as our identification strategy to estimate the impact of online accessibility on innovational strength. Section 4 discusses our main results and some robustness checks. Section 5 studies the effect of online accessibility on an alternative output-based measure of creativity, i.e., the number of fields and subfields an article contributes to. Section 6 concludes.

## 2 Measuring innovational strength

In the spirit of the epigraph by Henri Poincaré, we surmise that articles drawing from ideas which are "from domains far apart" are more innovative. We interpret this distance in a rather literal sense as path length between articles in the network of citations. Thus, we propose measuring the degree of innovation with which an article uses the existing literature by considering the distribution of pairwise shortest back-in-time citation paths between an article's references. Our approach builds on other research on originality, innovation and creativity in citing behavior. Hall *et al.* (2001) measure a patent's originality by computing a concentration index of cited patent classes.[3] Uzzi *et al.* (2013) consider, as we do, pairwise combinations of references; and they compute the likelihood of the pair of references' journals to be co-cited.[4] In contrast to these measures, the one we propose is (i) constructed at the level of the cited article rather than the cited 'class' or the cited journal, (ii) exploits the complete structure of the citation network, and (iii) reflects (in a stylized manner) the notion of combining two references for the first time. These features are important in our context of recombinant growth, as we wish to establish whether an article is joining strands of the literature that up to that point in time were unconnected. The proposed measure gives a precise quantification of how well connected an article's cited literature is at that point in time. Once an article joins two disparate strands of the literature by citing articles from both strands, this article serves as a bridge between the literatures. This article then permanently reduces the geodesic distances between the joined literatures for future articles.[5]

[3] See also Trajtenberg *et al.* (1992).

[4] I.e., the probability that a reference pair contains those two journals relative to a random network where articles display the same number of references and citations as in the actual network.

[5] Other measures that have been used in the literature —as discussed above— lack this feature. An article exactly copying an earlier highly creative or original article could obtain the same originality score.

Calculating our measure of innovational strength requires to construct the entire network of references such that back-in-time citation paths can be identified. The shortest paths or *geodesics* provide essential information about the structure of networks; many measures of network connectivity and centrality are characterized by functions of geodesics (Jackson, 2008). The analysis of geodesics in academic citation networks dates back at least to de Solla Price (1965). By focusing on the distribution of geodesics of an article's references, we capture a local connectivity specific to citation networks which, to the best of our knowledge, has not been explored in the literature so far.

### 2.1 The citation network

The citation network we consider is the one spanned by all articles published between 1955 and 2009 in 50 selected core journals of economic research. The list of journals includes the top five general interest journals,[6] the main top field and second tier general interest journals (as well as their historical predecessors). Table A.14 in the Appendix provides an alphabetic list of the journals.[7] Using eigenfactor.org's list of over 200 economics journals, we found that our list has an eigenfactor score of around 0.75 for the year 1995. I.e., randomly traversing the citation network of all economics journals, the list's 50 journals are selected in about 3 out of 4 times. Thus, our measure will reflect how articles cite and use this core economic literature; and, in turn, the analysis in the following sections will study how the digitization of this core literature affected the innovational strength of follow-on research. While digitization may also have impacted the use of more peripheral economic literature and the degree of references' interdisciplinarity, our approach does not capture these margins of the online accessibility effect.[8]

We downloaded from Thomson Reuters' Web of Science the bibliographic record of all items published between 1955 and 2009 in the 50 core economics journals.[9] The sample does not only include articles but also notes, letters, book reviews etc., which gives rise to a total of 129,145 items. The list of cited references is part of an item's bibliographic record. To construct the

[6] American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics, and Review of Economic Studies.

[7] The set of journals includes all journals considered in the standard Tilburg ranking, as well as the list considered in Palacios-Huerta & Volij (2004). Furthermore, it includes all core journals in Conroy *et al.* (1995), all journals used in Kalaitzidakis *et al.* (2003), as well as all top 20 journals in Combes & Linnemer (2010). The list is comparable to Depken & Ward's (2009) who include 79 economics journals. In contrast, McCabe & Snyder (2015) also consider business journals. Because our journal list includes 3 predecessors, only 47 of the selected journals publish nowadays.

[8] Our identification strategy (described in Section 3), however, takes such effects into account and is robust against them.

[9] The year 1955 is the earliest year available on Web of Science. Ten journals of our list of 50 were founded before 1955, three of them in the late 19th century (Economic Journal, Journal of Political Economy, and Quarterly Journal of Economics).

citation network, references were matched back to the published items. On average, we were able to match 36 percent of all references, and 44 percent of references in articles of our main sample years 1991-2009. Unmatched references may refer to publications prior to 1955 or to publications in books, working papers or journals which are not included in our sample. Finally, we calculated the shortest back-in-time connection within the citation network for all binary pairs of (identified) references.

### 2.2 Distribution of geodesics: an example

As an example, the calculation of the geodesics are illustrated in Figure 1 for a papers-and-proceedings article written by John Cochrane and Monica Piazzesi and published in 2002.[10] In Panel (a) of Figure 1, this article is visualized by a red node. The selected article cites seven references. Within our sample we can identify four of them and these items are depicted as blue nodes.[11] The four identified references give rise to six bilateral connections (unordered pairs) among them. We then calculate for each of the bilateral links the back-in-time geodesic within the entire citation network spanned by the sample of over 100,000 items. In the example of Figure 1, one of the references, *Rudebusch (1998)*, cites another one directly —*Christiano et al. (1996)*— which implies a geodesic distance of one (Panel b). Moreover, *Christiano et al. (1996)* is linked to another reference, *Cochrane (1989)*, via two connections; as is the case for *Rudebusch (1998)* and *Clarida et al. (2000)*. Panel (c) plots these geodesics of order two. The shortest connection from *Clarida et al. (2000)* to *Cochrane (1989)*, and to *Christiano et al. (1996)*; as well as between *Rudebusch (1998)* and *Cochrane (1989)*, is given by three steps (Panel d). We determine the geodesics iteratively up to a length of 3, giving rise to a probability mass function over four categories, with the last category comprising geodesic distances strictly larger than three. We denote these shares of references' pairwise geodesics of order one, two, three and higher than three for article $i$ by $S_{i,1}$, $S_{i,2}$, $S_{i,3}$ and $S_{i,>3}$. In the example of Figure 1, we have $S_1 = \frac{1}{6}$, $S_2 = \frac{1}{3}$, $S_3 = \frac{1}{2}$, and $S_{>3} = 0$, respectively.

This simple example, chosen for illustration purposes, is not typical for the dataset. The median article has 10 identified references and thus its references' citation network comprises 45 bilateral distances. Over the period 1991-2009 which we will use for our analysis of online accessibility, the average shares of the different geodesics are 25, 27, 20, and 28 percent (see Table A.1).

The way geodesic distances are constructed may generate an inherent time trend, since the citation

[10] "The Fed and Interest Rates: A High-Frequency Identification," *American Economic Review, Papers and Proceedings*, May 2002, Volume 92, Issue 2, pp. 90-95.

[11] In the following we abbreviate all the sources by authors and publication date in italic without specifying the entire reference. For the exact reference of the citations we refer the reader to the paper by John Cochrane and Monica Piazzesi.

network is more comprehensive for later years where our dataset covers more back volumes. This tends to make the share of geodesic distances higher than three falling over time. Therefore, it is essential that empirical analyses involving geodesics account for such time effects. In our regression framework we do so in a flexible manner by using publication year fixed effects.

— Figure 1 about here —

### 2.3 Geodesics between an article's references predict its long-term citations

In terms of a recombinant growth model of academic research, higher shares of long geodesics represent longer distances between the ideas an article builds on. As a rough validation check of our measure, we study whether articles whose references exhibit longer geodesics —and are thus expected to be more innovative— receive more citations.

The outcome variable is the number of citations an article $i$ received $K$ years after it has been published. Since we want to link citations to academic innovation, we focus on large $K$: innovative articles are those having a long-lasting impact on subsequent research.[12] The key explanatory variable is the share of geodesic distances of order three and higher ($S_{\geq 3}$). The results of Section 4 show that the main effect of online accessibility is to shift mass into this part of the distribution of geodesics. The empirical model we estimate is

$$Citations_{i,K} = \exp(\alpha\, S_{i,\geq 3} + \mathbf{X}_i'\beta + \mu_v)\,\eta_i, \tag{1}$$

where $Citations_{i,K}$ is the number of citations to article $i$, $K$ years after its publication, $\mathbf{X}_i$ is a vector of covariates and $\mu_v$ is a volume fixed effect. In the following we use the term 'volume' and the index $v = \tilde{v}\,(j,\tau)$ for all issues of a journal $j$ published in the same calendar year $\tau$. Finally, $\eta_i$ ($\geq 0$) is a unit-mean error term. We are interested in the sign and magnitude of $\alpha$.

The top panel in Table 1 displays estimates of (1) from a Poisson pseudo-maximum likelihood regression of citations after 20, 30 and 40 years on the share of geodesics greater or equal than three. To compute the citations for such long periods of time, we need to use our full network of citations back to 1955. We use a fixed effects estimator that accounts for volume- (or journal-year-) specific heterogeneity and we include controls for 'network variables'. These are an article's number of references (in the network), its share of self-references, the average number of its references' citations and of its references' references. The network variables ensure that the effect is based on differences between articles whose reference network's overall degree of connectivity is comparable. The volume fixed effects account for different citations patterns across journals

[12] Long-term citations reflect the social return of research, which may not necessarily be aligned with private returns of individual researchers (see Mandler, 2015).

(and thus, roughly, across fields) and time. As mentioned, controlling for time effects also serves the important purpose of holding the depth of the citation network constant.[13]

The bottom panel in Table 1 includes further controls: an indicator for whether the article appeared in a paper-and-proceedings issue, the number of authors, the number of pages, the total number of references (including references not in the network), and the number of journals referenced.

The Poisson regression is based on an exponential mean function so that estimated coefficients have the interpretation of semi-elasticities. For instance, consider the predicted change in citations for a unit change in the share of geodesics larger or equal to three. This change corresponds to the difference between articles in the 1st and 9th decile in the distribution of the share of geodesics. The results in Panel II indicate that *ceteris paribus* the article with the longer geodesics will receive about 11.5 percent more citations after 20 years. This advantage grows over time: after 30 years, the article is predicted to have about 31.7 percent more citations; and after 40 years, about 70 percent more. For shorter periods the effects were smaller and insignificant.[14] Thus, the results point to the distribution of geodesics being related to the long-term success of publications.

— Table 1 about here —

In Table A.2 in the Appendix, we also look at the articles in the far right tail of the distribution of citations. Using a fixed effects logit model, we predict the probability that an article is in the 95th percentile of citations among all articles published in the same year. Our results show that the odds of being in the 'top 5 percent' 20 years after publication is about 28.8 percent higher for an article with all its references' geodesics larger or equal than three than for an article with all geodesics of order one and two. 30 and 40 years after publication the effect is amplified to an increase in the odds of 40.8 and 139.3 percent, respectively.

Finally, we also looked at some anecdotal evidence stemming from the highest cited papers in

[13] Controlling for journal effects avoids, for instance, that field-specific citation norms confound the effect. An argument against using journal fixed effects is that the journal an article is published in is to some extent endogenous to its innovativeness. According to this view, journals should rather be viewed as mediating the effect. However, results without controlling for journal fixed effects (not shown) yielded qualitatively similar results to the ones shown in Tables 1 and A.2 in the Appendix.

[14] As can be seen from the changing number of observations, the sample is not kept constant across estimations with different dependent variables (different $K$). Thus, the models for citations after 40 years are estimated from older publications than those for citations after 30 years. This might rise the concern whether the growing effect is driven by the changing sample. However, re-estimating the models for citations after 30 and 20 years with the sample used for citations after 40 years gives the same increasing pattern, and so does re-estimating the model for citations after 20 years with the sample for citations after 30 years. The effects tend to be larger than with the sample for citations after 20 years.

our data. Only 14 articles amass over 1,000 citations from other articles in the network 30 years after publication. That 9 of these articles' authors would later go on to be awarded the Nobel Memorial Prize in Economics speaks of their degree of innovation.[15] While the mean of the share of references' geodesics that is three or longer equals 50 percent in the whole network and 65 for articles observed 30 years after publication, 11 of the 14 articles had shares larger than 65 percent. Indeed, the mean share for these articles is about 80 percent.

As a whole, the results in this section suggest that the distribution of references' geodesics is a measure which captures important aspects of articles' success in terms of citations. Although citations are subject to noise, the long-term perspective offered here should mitigate this problem. At the same time, the long-run is also amenable to the interpretation of citations as a reflection of academic innovation.

# 3 Identification strategy

## 3.1 Measuring online accessibility

To measure historical online accessibility of the literature, we combine data from Fulltext Sources Online (FSO) and JSTOR. The volume index $v = \tilde{v}(j, \tau)$ again refers to all issues of a journal $j$ published in the same calendar year $\tau$. The FSO data contains information about the online accessibility of the different journal-volumes at the journals' own webpage as well as all major platforms (such as, e.g., EBSCOhost, LexisNexis, ScienceDirect, or WilsonWeb). In the FSO data, we observe biannually for the years 1998-2009, for each volume $v$, whether it was accessible online on the different major platforms. An important exception, however, is the platform JSTOR, which has not been covered by FSO before the year 2009. Because JSTOR was (and still is) a very important provider of online access, we augment the information from FSO with data directly obtained from JSTOR. We combine the FSO and JSTOR data into an indicator of online accessibility, denoted by $a_{v,t}$. This indicator equals one if volume $v$ was accessible online on at least one platform (or the journal's own webpage) during the year $t$, and zero otherwise.[16]

Online accessibility has been stratified by the different online platforms at the journal-volume level. This variation, both over time and across volumes, is captured by our indicator of online

[15] The Prize was awarded to the authors on average about 25 years after publication of their respective (often cited) articles.

[16] We assume that before 1998, no volumes were accessible on platforms covered by the FSO data. This is reasonable, since only about 2 percent of volumes were online on platforms other than JSTOR in 1998, and these accessible volumes were mainly the contemporaneous ones. For the historically most important platform of online access, JSTOR, we do have the data about the accessibility of journals even prior to 1998.

access $a_{v,t}$. Figure 2 shows for three selected journals the historical online accessibility of all the volumes. The gray areas in Figure 2 (a) show for each point in time (x-axis) which of the publication years (y-axis) of selected journals were accessible online. Typically, online accessibility of a journal started by making recent publications accessible first, and then adding more and more older volumes. For instance in the case of the *Journal of Financial Economics*, the most recent volume went online first in the year 2000. In some instances, as for the *Review of Financial Studies*, the most recent volumes were never accessible online —a case of a so-called moving wall restriction. Across journals, there is considerable variation in the year the first volume was put online and how long the transition time lasted until all the old volumes were scanned. For example in the case of *Games and Economic Behavior* the transition to full coverage took 6 years. Figure 3 (a) plots for each year the average share of existing volumes in the 50 considered journals that were accessible online. Moreover, we also show for each year selected percentiles of the distribution across journals. Overall, the variation of the historical accessibility of journal-volumes is large, even in a group of journals of comparable quality or in the same subfield. A general observation is, however, that journals published by Elsevier —which are not covered by JSTOR— were lagging behind.[17] Online accessibility of economics journals started in 1997 on the JSTOR platform. Since then, the back volumes of the different journals were gradually scanned and uploaded, and in 2009 virtually all publications were available online. Hence, the period considered covers some years of the pre-internet era as well as the entire transition to full coverage. Until the turn of the millennium online access was dominated by JSTOR. Later, other platforms caught up. A large amount of back volumes of Elsevier journals was made accessible in the year 2005.

— Figure 2 about here —

Importantly, however, our identification strategy relies on the cross-sectional variation rather than on variation over time. For this we construct an article-specific variable called *Share online*, that measures the fraction of the article's relevant literature that was accessible online in a year $t$. In the following, an article is indexed by $i$. We define the set of journals an article $i$ cites as this article's relevant literature.[18] More formally, suppose $\boldsymbol{V}_{i,t}$ denotes the set of volumes published in the years $\tau \leq t$ in a journal $j$ which is cited by article $i$. *Share online* measures the share of all

[17] Among the three journals shown in Figure 2 only the *Review of Financial Studies* is not published by Elsevier and contained in JSTOR.

[18] In a robustness section we show that our results are robust to other reasonable definitions of relevant literature. Evans (2008), whose units of observation are journal-years, defines the relevant past literature as the journal where an article is published in. This approach seems unsuitable for economics. In our data, on average only 7 percent of citations refer to the same journal where the article is published. Even for the journal where this ratio is highest over the whole period —the Journal of Finance— it is only 20 percent.

volumes of cited journals that were accessible online on at least one platform:

$$\textit{Share online}_{i,t} = \frac{\sum_{v \in \boldsymbol{V}_{i,t}} a_{v,t}}{\sum_{v \in \boldsymbol{V}_{i,t}}}. \tag{2}$$

Our identification strategy will exploit this variation in the online accessibility of the relevant literature across articles published in the same journal-volume $v$. The measure of online accessibility is article-specific, because the relevant literature depends on the set of cited journals (and the fractions of accessible back volumes differs across journals). As can be seen from Figure 3 (b), in the year 1999, a large fraction of back volumes of the *Review of Financial Studies* had already been made accessible online. In contrast, however, in the case of the *Journal of Financial Economics* only a very small fraction of the volumes were accessible online until the year 2004, and for *Games and Economic Behavior* the fraction of accessible volumes was gradually increasing until 2006. Now suppose an article cites the *Review of Financial Studies* and *Games and Economic Behavior*. In the year 2003, our measure of *Share online* for this article is then 0.64 (formally this is a weighted average of the two journal's share online, where the weights are given by the number of published volumes in each journal). Now suppose another article cites instead the *Review of Financial Studies* and the *Journal of Financial Economics*. In the same year 2004, this article's measure of *Share online* is only 0.37. The intersecting lines in Figure 3 (b) also make it clear that the ranking of *Share online* of the same two articles would be opposite in the year 2004 or after. Our identification strategy exploits precisely this variation across journals and time to estimate the effect of online accessibility.

In addition to this main empirical approach, Section 4.6 explores an alternative identification strategy based on citation-level data (citing article–cited article pair), which circumvents the need of defining a relevant literature for the citing article. The results from both approaches are very similar, but ultimately we favor the present approach because its vantage point is more directly that of the citing literature, which is our main interest.

### 3.2 Econometric specification

In our analysis, an observation is an article $i$ published in journal $j$ and year $\tau$, i.e., in volume $v = \tilde{v}(\tau, j)$. To identify the effect of online accessibility of the literature, we run the following type of OLS regressions:

$$S_{i,l} = \alpha \times \textit{Share online}_{i,t=\tau-2} + \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{\mu}_v + \epsilon_i, \quad l = 1, 2, 3, >3; \tag{3}$$

where the dependent variable $S_{i,l}$ is article $i$'s fraction of bilateral geodesics of order $l$ (=1,2,3,>3) between its references, and $\epsilon_i$ is an error term. The symbol $\boldsymbol{\mu}_v$ stands for a full set of volume (i.e., journal-year) fixed effects and $\boldsymbol{X}_i$ represents additional article-specific control variables. By

setting $t = \tau - 2$ we evaluate our measure of online accessibility two years prior to the publication year of the citing article. This time lag takes into account that the publication process takes some time and implicitly assumes that the authors did their literature search two years prior to publication.[19] We are interested in the coefficient $\alpha$, that measures the effect of a marginal increase in the share of literature online on the distribution of the geodesics. Equation (3) can be estimated conveniently using the OLS within-estimator.

The default behavioral model behind this identification strategy is that an author facing zero online accessibility searches the relevant literature in print, for instance by browsing his library's collection of volumes aided by keyword or abstract indexation systems. In contrast, another author whose relevant literature is partially accessible online will browse this electronic literature by using internet tools, while still using the same methods as the previous author for the literature available only in print. In such a case, the use of internet literature browsing and searching tools coincides exactly with our online treatment variable. In practice, some deviations from such behavior are likely. For instance, very low levels of *Share online* might not induce researchers to search online; and, conversely, researchers whose literature is almost entirely online might neglect the few remaining print-only volumes. However, studies on researchers' literature searching behavior suggest that the joint use of print and electronic resources (with declining use of print) was typical for researchers during the transition to full electronic access (Tenopir *et al.*, 2003; Boyce *et al.*, 2004), so that *Share online* should be a reasonable approximation to researchers' behavior.[20] In our main analysis the units of observation are articles (i.e. 'ideas') rather than their authors. In a robustness section, however, we document the same qualitative results at the co-author group level.

A qualification needs to be made at this point. While we compute an article's share of relevant literature which was *accessible* online, we do not observe whether the article's author(s) effectively *used* the internet to search for related literature. Thus, online accessibility effects should be understood as intention-to-treat effects of online access.

By estimating the effect of online accessibility within a journal-volume, specification (3) controls flexibly for time trends. Hence, our specification will be able to disentangle the effect of online accessibility from any other ongoing trends.[21] The journal-volume fixed effect specification also controls for the exogenous heterogeneity in the innovational degree across journals, which other-

[19] In Section 4 we show that our results are robust to assuming a publication lag of 0–4 years.

[20] An alternative interpretation of the online treatment variable comes from a more stylized behavioral model where there exist only two types of researchers: one group using print literature only, the other group relying exclusively on online literature. Then, *Share online* can be interpreted as the probability that the article's author is an online researcher.

[21] See also McCabe & Snyder (2015) who illustrate, in a slightly different context, the empirical importance of flexible controls for time trends.

wise could have been a source of spurious correlation. At the same time one needs to keep in mind that the volume fixed effects also excludes some margins which might well be affected by the online accessibility of the literature. For instance, if a high online accessibility of a specific literature makes it more likely that papers in this literature end up being published in general interest journals with a higher (average) innovational strength, this will not be reflected by our estimates.[22]

Although both innovational strength and online accessibility are constructed from an article's reference list there is no mechanic reverse effect from citing behavior on the measure of online accessibility. Since our measure of online accessibility is the *average* over all past volumes of cited journals, the number of journals cited or the publication years of the cited references has no mechanic effect on *Share online* (and we show as a robustness check that controlling for these variables does not change our results). It is, however, a fact that none of the main general interest journals is printed by Elsevier. Thus, a related concern might be that our results are partially driven by the quality of the cited journals. To exclude this possibility, we always include (as part of our network control variables) the average number of citations received by an article's references in $\boldsymbol{X}_i$. In that way, we basically compare two articles, each of which cites a set of references which are equally well-cited but differ in their online accessibility. In addition, we show that our results are robust to the inclusion of the share of references going to top five journals in the regressions. Finally, we show that qualitatively similar effects are obtained if we control for author fixed effects.

# 4 The effect of online accessibility on academic innovation

## 4.1 Online accessibility shifted references' geodesics towards longer distances

Table 2 reports the coefficient estimates of *Share online* for the baseline model (3). The three columns represent regressions where the dependent variable is the fraction of the geodesics between an article's references of length one, two and three (labeled $S_l$ for $l = 1, 2, 3$). Results for a fourth regression on $S_{>3}$ have been omitted; the coefficients would be redundant as they are numerically the same as minus the sum of the coefficients across the three columns, a result stemming from $S_{>3} = 1 - S_{\leq 3}$ and that changes in predicted fractions necessarily add up to zero. In all regressions, the OLS within-estimator is calculated over journal-years, of which there are 859 unique groups.

— Table 2 about here —

[22] If we only control for year fixed effects, our estimates get indeed larger in magnitude.

The yearly average of *Share online* varies from zero in 1991 to one in 2009, and thus a convenient interpretation of the coefficients is as estimates of the total change in the average fraction of geodesics associated with moving from a world without any online access to a world which provides full access to all volumes of the 50 journals. The results in Panel I indicate that online accessibility has shifted the probability mass in the distribution of references' geodesics from distances of order 1 and 2, which together make about half of the mass in the data, towards longer distances. For instance, the full effect of *Share online* decreases the fraction of geodesics of order 1 and 2 by 4.07 and 3.19 percentage points, respectively. The effects are statistically significant and of substantial magnitude, ranging from 11 to 21 percent of the dependent variables' means or from 18 to 27 percent of their standard deviations.

As discussed previously, the use of the within estimator controls for any confounding journal-year-specific characteristics. Regressions in Panel I of Table 2 also control for article-level characteristics of an article's network of references by including the number of references (in our network of 50 journals), the fraction of self-references (reference to an article with at least one joint author), the average number of citations made to the references, and the average number of references' references. In this way the online accessibility effect is estimated conditional on reference networks having the same degree of connectivity. Controlling for the average number of received citations of the references also controls for the quality of the references. Estimates in Panel II were obtained adding further article-level control variables to the specification: an indicator of whether the article appeared in a paper-and-proceedings issue, and the numbers of authors, pages, references (inside as well as outside of our journal network) and journals referenced. The effects are somewhat smaller than before, but remain large and statistically significant. The decrease in effect size is mainly the result of controlling for number of pages and number of different journals referenced, two variables mediating the effect of online accessibility. Whether the additional variables in Panel II are part of the causal effect and should not be controlled for is to a large extent a matter of taste and interpretation; preferring to err on the conservative side, we will adopt the specification in Panel II for all further regressions. Table A.3 in the Appendix contains the full results showing coefficient estimates for all control variables.

### 4.2 Robustness of the findings

An important first robustness check for our results relates to the appropriate lag of the treatment. The time when an article's references were collected is unknown and has to be inferred from the date of publication. In addition, there are also bound to be differences in length of the publication process across articles. In building our treatment variable we made the informed guess that a

good approximation is the online accessibility authors faced two years prior to publication.[23] Table A.4 in the Appendix explores alternative lags of zero, one, three and four years. Given the heterogeneity in publication process length, it would be worrisome to find that the results in Table 2 hold only under the one year lag. However, the results remain qualitatively the same for all lags explored.

Next, we set out to assess the robustness of our treatment by exploring other ways of capturing online accessibility. Implicitly, the treatment *Share online* gives more weight to long-standing journals (with many volumes) because the percentage is calculated over the sum of all volumes of cited journals. An alternative which weights journals equally is to construct the treatment as the share online in the average journal cited. Similarly, treatment can be defined as the percent of an article's references that was online two years prior to publication. This weights the journals by their share in the reference list. Finally, instead of focusing on percentages, treatments can also be constructed based on the *absolute number* of volumes online (an approach related to Evans, 2008). Table A.5 in the Appendix documents that the baseline results from Table 2 remain valid for any of these alternative treatments.[24]

We now turn to assessing robustness with respect to the data sources of online access. Our measure of online accessibility combines information obtained directly from JSTOR with information collected by FSO. Detailedness and quality of these two sources varies. Whereas FSO collects its data twice a year, JSTOR's database is very precise.[25] To make sure that such differences between data sources are not influencing our results, we constructed two treatments: one taking into account access provided by JSTOR only, the second measuring access on the remaining online platforms covered by the FSO data. The results, displayed in Table 3, show that disaggregating the treatment by data source delivers estimates similar to the aggregate treatment in the baseline specification. Moreover, McCabe & Snyder (2015) have hand-collected information on online access for a subset of our journals in the sample. In Table A.6 we use their online access data to construct our treatment variable.[26] As before, the results show that there was a shift from geodesics of order 1 and 2 to longer distances.

— Table 3 about here —

[23] Evans (2008) uses a one-year lag, while Depken & Ward (2009) and McCabe & Snyder (2015) use a lag of zero years.

[24] With some of the alternative treatments a change of one unit in the treatment does not have the interpretation of a switch to full online access anymore and so coefficient sizes are not directly comparable. Therefore, Table A.5 reports effect sizes for a switch to full online access.

[25] In fact, we know from JSTOR for each journal issue the exact date of first user access.

[26] The McCabe-Snyder data covers the period up to 2005 and there is no information on online access to 10 of the 50 journals included in our list.

The next issue we explore is a refinement of the fixed effects. In a first step, we treated papers-and-proceedings issues of a journal as a separate journal. Since most journals publish such issues, the number of panel units for these regressions increased to 1,456. Taking this approach one step further, we defined a separate fixed effect for every single issue published in every journal in the period. This gives over 4,800 fixed effects. As the estimates in Table A.7 in the Appendix show, these specifications confirm the baseline results.

### 4.3 Heterogeneity over time and across journals

Time is a potential source of heterogeneity in the effect. While there are many potential factors with a time trend, one of them has been highlighted in the literature as particularly relevant for online accessibility: institutional subscription to platforms providing online contents of economics journals (i.e., effective online access). Depken & Ward (2009) and McCabe & Snyder (2015) document that the number of institutions subscribing to JSTOR (and to Elsevier's online contents) increased almost linearly in the period considered (cf. Depken & Ward, 2009, Fig. 1, McCabe & Snyder, 2015, Fig. 7). Table 4 shows estimation results for a specification which adds an interaction of the treatment with a linear time trend, which is bound to capture this effect of increasing online access. The time trend was normalized to zero in 1997, so that the coefficient on *Share online* gives the effect in that year. For the following years, the coefficient on the interaction gives the yearly change in the effect. However, the estimated coefficients on the interactions are small and not statistically significant, so that the absence of a time trend cannot be rejected.

— Table 4 about here —

The effect of online accessibility found in our baseline regressions could also vary across different journals. While for the average journal the effect on the distribution of geodesics is positive, it could be that this aggregation masks negative effects for some journals. To explore this issue we estimated a specification with interactions for three classes of journals: the top five journals, other general interest journals, and field journals (see Table 5).[27] We find the same kinds of effects as in the baseline regressions for every journal category: a shift from geodesics 1 and 2 to larger distances. Overall, however, the shift seems to have been specially strong among articles published by the top five journals. With about 10 percentage points of the probability mass being transferred to distances of order 3 or larger, the effect is about twice as large as in other journals.

— Table 5 about here —

[27] Note that uninteracted level effects of journal classes are subsumed in the journal-year fixed effects.

### 4.4 A closer look at the composition of references: journals and publication years

One way in which online accessibility may have influenced the distribution of geodesic shares is by reducing the bibliographic importance of the journal an article appeared in. The correlation between reading a particular journal and contributing to it might have been weakened by the internet, leading to a more diverse pool of influences. Panel I in Table 6 addresses this issue by including an additional regressor for the percent of an article's references that is made to papers published in the same journal that the article appeared. As expected, the coefficients on this variable indicate that a higher share of references to the own journal goes in hand with smaller distances between references. The coefficients on *Share online* have the same pattern as in the baseline, but are somewhat attenuated. For instance, the probability mass shifted to geodesics of order 3 or higher is about 20 percent higher in the baseline results. This suggests that, indeed, a channel through which the online accessibility works is by reducing the share of references made to the own journal.

A second potential mediator of online accessibility is the share of references made to articles in top five journals, an issue explored in Panel II of Table 6. Top five journals might have benefited from online accessibility by ranking effects of online search engines, or —since their online presence was relatively prominent in the early phase of digitization— by first-mover-advantage effects. The estimates indicate that articles referencing higher shares of top five papers draw from more tightly connected literatures. Holding these shares constant, the online accessibility effect is stronger, with the shift to the longer half of the geodesics increased by over 30 percent compared to the baseline results. Taken together, the results from Table 6 illustrate that, although the overall effect on the distribution of geodesics is positive, the online accessibility effect encompasses channels working in opposite directions.

— Table 6 about here —

Finally, we set out to quantify the importance of the age distribution of an article's references for the effect on the distribution of geodesics. The results from regressions including average citation lag (i.e., the difference in years between the article's publication year and that of its average reference) are shown in Table 7. The average citation lag has been used as the primary dependent variable in previous work analyzing the impact of online accessibility on academic research (Evans, 2008; Depken & Ward, 2009). Our coefficients of interest remain virtually unaffected in size and statistical significance when including citation lag, showing that our outcome captures a fundamentally different dimension of an article's references than its age distribution.[28]

[28] Table A.8 contains further regressions including the median and standard deviation of references' publication

— Table 7 about here —

### 4.5 Results at the author level

A more fundamental extension changes the panel dimension to a much less aggregated unit: the authors. While in our baseline regressions we exploit the variation in online accessibility between articles of a particular journal in a given year, a different source of variation comes from repeated publications of the same coauthor groups. Exploiting only the variation for a given coauthor-group changes the interpretation of the coefficients, as the online accessibility effect being estimated excludes channels which are part of the effect using the within journal-year variation. For instance, the availability of online literature may have an impact on the composition of the pool of authors, increasing the share of authors who are efficient users of online tools. In the estimation with journal-years fixed effects this margin is part of the causal effect as the pool of authors is not kept constant and changes with the spread of online accessibility. While ultimately we favor this approach, the specification with coauthor fixed effects provides an important alternative view which shows the effect of online accessibility for authors publishing repeatedly during this period.

— Table 8 about here —

We extracted author names from the EconLit database[29] and used unique coauthor-groups (including groups of size one, i.e., single authors) as the panel unit. There are $7,307$ such unique coauthor-groups who have published more than twice in our data. The total number of articles they have published is $21,767$. In the regression, we additionally include over 800 journal-year fixed effects and our list of control variables. The results, printed in Table 8, are substantially less precise. Given the substantial loss of degrees of freedom, this does not come as a surprise. It is the more remarkable, therefore, that the results in this table reveal the same patterns than those from the baseline regressions. To be sure, the coefficients are visibly attenuated compared to the baseline; still, we find that online accessibility transferred probability mass from the distribution of geodesics' lower end ($S_1$ and $S_2$) to its right tail.

### 4.6 An alternative identification strategy: Citation-level data

To conclude the section, we present results from a different empirical approach which uses information at the citation-pair level. Compared to our main identification strategy, there are two

years yielding very similar results.

[29] We used data from EconLit as we found it more consistent in the coding of author names than Thomson Reuters.

important advantages from using pair-level data: It avoids the need of specifying a relevant literature for the citing article, and it allows to control flexibly for characteristics of cited journal volumes. Let $i$ index citing articles and let $k$ index cited articles. Online availability, the treatment variable, is now defined by the indicator $Online_{ik,t=\tau-2}$, which is equal to one if in year $t$ the volume of the cited article $k$ was accessible online. As before, we evaluate $t$ with a two-year lag relative to $i$'s publication year $\tau$. The econometric model to be estimated is

$$S_{i,l} = \alpha \times Online_{ik,t=\tau-2} + \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{\mu}_{v(i)} + \boldsymbol{\nu}_{v(k)} + \epsilon_{ij}, \quad l = 1, 2, 3, >3. \tag{4}$$

The outcome variable is still the citing article's share of geodesics of different lengths, and we again control for citing-article characteristics $\boldsymbol{X}_i$ as well as for citing-article volume fixed effects, $\boldsymbol{\mu}_{v(i)}$. In addition, (4) includes a full set of cited-article volume fixed effects, $\boldsymbol{\nu}_{v(k)}$. This is possible because, by adopting a pair-level approach, we can compare the average innovational strength of citing articles of a given cited volume, before and after that cited volume was made accessible online.

— Table 9 about here —

Table 9 presents the regression results for this approach. The sample comprises 406,633 citations. Cited articles have been restricted to publication years 1985 or later, because early volumes are cited relatively infrequently in the data creating problems of collinearity and precision.[30] The coefficient $\alpha$ in (4) gives the effect of a single cited articles' online accessibility on a citing article's shares of geodesic distances. To make the results comparable in magnitude to those of the previous approach, Table 9 shows the effect of full online accessibility for an average citing article (that is, $\alpha$ multiplied by 35.95 which is the average number of cited articles). The results confirm our previous findings that online accessibility shifts the probability mass from short geodesics of order 1 and 2 to long geodesics of order 3 and larger than 3. The effect sizes are about twice as large as those found with the previous approach. In addition, Table 10 displays results for the additional specifications from Sections 4.3 and 4.4. For compactness, the dependent variable is $S_{\geq 3}$, the share of geodesics of order 3 and larger. Again, the results confirm our previous findings. In each case, the effects showing an increase in the shares of longer geodesics are large and statistically significant.

— Table 10 about here —

Taken together, these results show that the findings from our main identification strategy are

[30] Table A.9 in the Appendix gives the results for the sample with cited articles dating back to 1955 (sample size: 522,692). The results are qualitatively similar although less precise.

neither driven by the definition of the relevant literature nor by possible endogeneity of the treatment stemming from correlation with unobserved characteristics of cited journal volumes.

# 5 Online accessibility increased the number of articles' unique JEL codes

The previous section analyzed the effect of online accessibility on the intellectual input into the academic production function. We document that because of online accessibility authors read more heterogeneous strands of the literature and build on ideas which have not been combined before. Our measure of innovational strength is clearly rationalized by a recombinant growth theory. Furthermore, we illustrated in Section 2.3, that this measure of innovativeness possesses predictive power for the number of future received citations. Nevertheless, it is the aim of this section to see whether qualitatively similar effects can be found on alternative outcome measures as well. For this, we use an article's number of unique JEL codes as a measure of the breadth of an article's content, and investigate the impact online accessibility had on it.

## 5.1 The JEL classification system

EconLit, the American Economic Association's electronic bibliography, assigns up to six three-digit JEL classification codes to publications, and we downloaded this information from the EconLit webpage.[31] The JEL classification indexes the contents of an article, describing which fields and subfields it falls into. The American Economic Association introduced this classification system in 1991, and consequently we observe these classification codes from then onwards.[32] The first digit of a JEL code is a letter which divides economics into twenty main fields, such as "labor and demographic economics" or "industrial organization". While half the articles fall into exactly one field according to the one-digit definition, about 37 percent contribute to two fields, and somewhat over 10 percent have three one-digit JEL codes. There are large differences between journals. For instance, while the average article in *Econometrica* has about 1.1 one-digit JEL codes, the *Journal of Development Economics*' average article has about 2.3. The variation within journals is even larger (see also Table A.1). Using again journal-year fixed effects, this variation within a journal is the one we exploited. The last two digits classify the twenty fields

[31] These JEL codes are assigned by a team of economists at EconLit. Thus, they can and do differ from JEL codes declared by authors or reported by journals.

[32] See Pencavel (1991), the editor's note with which the *Journal of Economic Literature* introduced the new system. The JEL codes replaced an earlier, narrower classification system.

into narrower sub- and subsubfields resulting in a very subtle measure of article breadth.[33] The median article has two three-digit JEL codes, while about one third of articles have more than two. In our analysis we consider the number of distinct one-, two- and three-digit JEL codes each as a separate dependent variable.[34]

Figure A.1 in the Appendix plots the average number of one-, two- and three-digit JEL codes assigned to an article. For the first years in our sample, the number of assigned codes is constant. From 1995 onwards, it rises for all JEL code digits. Since the number of codes are not perfectly comparable between different years, we do not want to overstate these dynamics. For instance, some additional JEL codes were added after 1991; also, the production process set an upper bound of five codes assigned to an article until the mid 1990s; and, finally, the assigning process might have changed over time: e.g., the dint in the number of codes in 2006 might be explained by a change of EconLit's managing director. For all these reasons, it is indispensable to control for year fixed effects to disentangle the effect of online accessibility from other ongoing trends.[35]

### 5.2 Estimates of the effect of online accessibility on the number of JEL codes

The top panel in Table 11 contains results for regressions of an article's number of 1-, 2- and 3-digit JEL codes on *Share online* and journal-year fixed effects. The bottom panel includes all the additional control variables used in the full baseline specification of the regressions on geodesic distances. The results in both panels indicate that online accessibility had a positive, statistically significant impact on the number of fields and subfields an article contributes to. The effect of a change from zero to complete online accessibility amounts to about 30 percent of the variables' standard deviations in the regressions from Panel I; or around 25 percent of the variables standard deviations in the specification with further controls. Relating the results from Panel II to the raw time trend of Figure A.1, they correspond to about 25 to 32 percent of the change in the average number of JEL codes between the first year of our data, 1991, and the last one, 2009.

— Table 11 about here —

[33] JEL codes constitute a unique and precise categorization of articles' contents beyond its main field, which other similarly structured applications, such as patent citations, lack. In such datasets the intellectual content of a patent is limited to one "patent class" only.

[34] JEL codes have been the subject of some descriptive work which used them to characterize the evolution of economic fields or subfields over time (Kim *et al.*, 2006; Kelly & Bruestle, 2011). Previous literature using JEL codes in regression analysis has included them as control variables for the specific fields (e.g., Formby *et al.*, 1993; Axarloglou & Theoharakis, 2003; Boschini & Sjögren, 2007).

[35] The assignment of JEL codes is consistent within a given year (personal communication with the Managing Director of EconLit).

While we do not report further results here, we found that the effect of *Share online* on the JEL codes variables has the same robustness than the effect on geodesics across a number of specifications and extensions. In addition, similar effects are found when using the pair-level identification strategy as in equation (4), which in addition controls for cited journal-volume fixed effects (Table A.11 in the Appendix). Overall, we interpret the results of Table 11 as evidence that *share online*'s effects extend beyond directing researchers' attention to more heterogeneous literature to also affecting the breadth of the research's contents.

## 6 Concluding remarks

Using a novel measure of innovational strength based on the geodesics between an article's references, this paper documented how the digitization of the scholarly literature led to more innovative economic research in the sense of promoting a more creative use of the existing core literature. As a result of online accessibility, the use of references that up to that point were less connected in the citation network became significantly more likely. In the past, such combinations of less connected literature strands have been associated with higher long-term citation counts. Using the number of unique JEL codes assigned by EconLit as a measure of content breadth, we found that online accessibility also led researchers to touch on more subjects in their writings.

Our measure of innovational strength or creativity could also, in some sense, be understood as a measure of diversity. Since in this reading our measure relates to the diversity of ideas *a single article* is based on, our analysis is one of *local* diversity, and complements the aggregate measures of diversity considered in the previous literature (Evans, 2008; McCabe & Snyder, 2015). It can well be that the local diversity increases (as our results suggest) at the same time as the number of overall cited articles decreases and the concentration of cited articles increases (as suggested by Evans, 2008). For instance, different fields of economics may get tighter connected, whereas in each field some 'superstars' emerge.[36] However, the results of McCabe & Snyder (2015) suggest that in the case of economics and business, online access did not skew the distribution of citations.

Finally, while we have mainly interpreted the increase in creativity of citations brought about by the digitization of the literature and the use of new searching tools in light of recombinant growth theory, other interpretations are possible; and while the effect found seems desirable in a recombinant growth framework, this need not be the case for other ways of interpreting the results. Under a Kuhnian outlook (Kuhn, 1962), for instance, higher levels of 'diversity' produced

[36] For instance, a growing theoretical literature suggests that as more articles become available, attention gets scarce in the knowledge accumulation process, with implications on the aggregate diversity (see, e.g., Franck, 1999; Klamer & Dalen, 2002; Falkinger, 2007a,b, 2008).

by online accessibility could represent unnecessary noise, artificially hampering the consolidation of focused fields within the discipline. Nevertheless, at the very least, the academic community rewards higher innovational strength, as captured by our measures, with more citations.

## References


[1] Acemoglu, Daron, 'Diversity and technological progress.' In *The Rate and Direction of Inventive Activity Revisited*: University of Chicago Press, pp. 319–356, *2011*.

[2] Axarloglou, Kostas & Vasilis Theoharakis, 'Diversity in economics: An analysis of journal quality perceptions.' *Journal of the European Economic Association*, **1** (6), pp. 1402–1423, *2003*.

[3] Boschini, Anne & Anna Sjögren, 'Is team formation gender neutral? Evidence from coauthorship patterns.' *Journal of Labor Economics*, **25** (2), pp. 325–365, *2007*.

[4] Boyce, Peter, Donald W King, Carol Montgomery, & Carol Tenopir, 'How electronic journals are changing patterns of use.' *The Serials Librarian*, **46** (1–2), pp. 121–141, *2004*.

[5] Combes, Pierre-Philippe & Laurent Linnemer, 'Inferring missing citations: A quantitative multi-criteria ranking of all journals in economics.' *Groupement de Rechereche en Economie Quantitative d'Aix Marseille (GREQAMJ), Working Paper No. 2010-28*, *2010*.

[6] Conroy, Michael E, Richard Dusansky, & Arne Kildegaard, 'The productivity of economics departments in the US: Publications in the core journals.' *Journal of Economic Literature*, **33** (4), pp. 1966–1971, *1995*.

[7] Depken, Craig A & Michael R Ward, 'Sited, sighted, and cited: The effect of JSTOR in economic research.' *SSRN Working Paper Series, Working Paper No. 1472063*, *2009*.

[8] Evans, James A, 'Electronic publication and the narrowing of science and scholarship.' *Science*, **321** (5887), pp. 395–399, *2008*.

[9] Falkinger, Josef, 'Attention economies.' *Journal of Economic Theory*, **133** (1), pp. 266–294, *2007a*.

[10] ——— , 'Distribution and Use of Knowledge Under the Laws of the Web.' *CESifo Working Paper, No. 2125*, *2007b*.

[11] ——— , 'Limited Attention as a Scarce Resource in Information-Rich Economies.' *Economic Journal*, **118** (532), pp. 1596–1620, *2008*.

[12] Formby, John P, William D Gunther, & Ryoichi Sakano, 'Entry level salaries of academic economists: Does gender or age matter?' *Economic Inquiry*, **31** (1), pp. 128–138, *1993*.

[13] Franck, Georg, 'Scientific Communication–A Vanity Fair?' *Science*, **286** (5437), pp. 53–55, *1999*.

[14] Ghiglino, Christian, 'Random walk to innovation: why productivity follows a power law.' *Journal of Economic Theory*, **147** (2), pp. 713–737, *2012*.

[15] Hall, Bronwyn H, Adam B Jaffe, & Manuel Trajtenberg, 'The NBER patent citation data file: Lessons, insights and methodological tools.' *National Bureau of Economic Research, Working Paper No.8498*, *2001*.

[16] Jackson, Matthew O, *Social and economic networks*: Princeton, NJ: Princeton University Press, *2008*.

[17] Kalaitzidakis, Pantelis, Theofanis P Mamuneas, & Thanasis Stengos, 'Rankings of academic journals and institutions in economics.' *Journal of the European Economic Association*, **1** (6), pp. 1346–1366, *2003*.

[18] Kelly, Michael A & Stephen Bruestle, 'Trend of subjects published in economics journals 1969–2007.' *Economic Inquiry*, **49** (3), pp. 658–673, *2011*.

[19] Kim, E Han, Adair Morse, & Luigi Zingales, 'What Has Mattered to Economics Since 1970.' *Journal of Economic Perspectives*, **20** (4), p. 1, *2006*.

[20] Klamer, Arjo & Hendrik P van Dalen, 'Attention and the art of scientific publishing.' *Journal of Economic Methodology*, **9** (3), pp. 289–315, *2002*.

[21] Kuhn, Thomas S, *The structure of scientific revolutions*: Chicago, IL: University of Chicago Press, *1962*.

[22] Mandler, Michael, 'The Benefits of Risky Science.' *Economic Journal (forthcoming; doi: 10.1111/ecoj.12324)*, *2015*.

[23] McCabe, Mark J & Christopher M Snyder, 'Does online availability increase citations? Theory and evidence from a panel of economics and business journals.' *Review of Economics and Statistics*, **97** (1), pp. 144–165, *2015*.

[24] Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, & Scott Stern, 'Of mice and academics: Examining the effect of openness on innovation.' *National Bureau of Economic Research, Working Paper No. 14819*, *2009*.

[25] Palacios-Huerta, Ignacio & Oscar Volij, 'The measurement of intellectual influence.' *Econometrica*, **72** (3), pp. 963–977, *2004*.

[26] Pencavel, John, 'Editor's note.' *Journal of Economic Literature*, **29** (1), p. v, *1991*.

[27] Poincaré, Henri, 'Mathematical creation.' *The Monist*, **20** (3), pp. 321–335, *1910*.

[28] de Solla Price, Derek J, 'Networks of scientific papers.' *Science*, **149** (3683), pp. 510–515, *1965*.

[29] Tenopir, Carol, B Hitchcock, & A Pillow, 'Use and users of electronic library resources: An overview and analysis of recent research studies. Council on Library and Information Resources.' *Washington, DC: Council on Library and Information Resources*, *2003*.

[30] Trajtenberg, Manuel, Rebecca Henderson, & Adam Jaffe, 'Ivory tower versus corporate lab: An empirical study of basic research and appropriability.' *National Bureau of Economic Research, Working Paper No. 4146*, *1992*.

[31] Uzzi, Brian, Satyam Mukherjee, Michael Stringer, & Ben Jones, 'Atypical combinations and scientific impact.' *Science*, **342** (6157), pp. 468–472, *2013*.

[32] Weitzman, Martin L, 'On diversity.' *Quarterly Journal of Economics*, **107** (2), pp. 363–405, *1992*.

[33] ——— , 'Hybridizing growth theory.' *American Economic Review*, **86** (2), pp. 207–212, *1996*.

[34] ——— , 'The Noah's ark problem.' *Econometrica*, **66** (6), pp. 1279–1298, *1998a*.

[35] ——— , 'Recombinant growth.' *Quarterly Journal of Economics*, **113** (2), pp. 331–360, *1998b*.

# Tables

Table 1: Estimates of models for an article's total number of citations (Poisson pseudo-likelihood regressions with journal-years fixed effects)

| Citations... | ...after 20 yrs. | ...after 30 yrs. | ...after 40 yrs. |
|---|---|---|---|
| I. *Regressions controlling for network variables* | | | |
| $S_{\geq 3}$ | $0.2844^{***}$ | $0.4371^{***}$ | $0.7586^{***}$ |
| | (0.0646) | (0.1356) | (0.1937) |
| N | 21,477 | 7,694 | 1,520 |
| $R^2$ | 0.3217 | 0.3026 | 0.3370 |
| II. *Regressions controlling for network and further variables* | | | |
| $S_{\geq 3}$ | $0.1146^{*}$ | $0.3168^{**}$ | $0.7007^{***}$ |
| | (0.0637) | (0.1300) | (0.2066) |
| N | 21,477 | 7,694 | 1,520 |
| $R^2$ | 0.3625 | 0.3303 | 0.3720 |

*Notes:* $^{*}$, $^{**}$ and $^{***}$ indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by fixed effects Poisson pseudo-likelihood estimation accounting for journal-year fixed effects. Robust standard errors clustered at journal-year level in parentheses. $R^2$ is McFadden's pseudo R-squared. The dependent variables are the total number of citations received by an article 20, 30 and 40 years after publication. The key explanatory variable, $S_{\geq 3}$, is the share of geodesics of order greater than two among geodesics between an article's references. Network control variables: number of references in data, percent self-references, average number of references' citations, average number of references' references. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced.

Table 2: Fixed effects regressions of share online on geodesics, N=45,553

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| I. *Regressions on share online and network variables* [a] | | | |
| Share online | -0.0407*** | -0.0319*** | 0.0409*** |
| | (0.0109) | (0.0098) | (0.0076) |
| $R^2$ | 0.0679 | 0.0685 | 0.0760 |
| II. *Regressions on share online and further variables* [b] | | | |
| Share online | -0.0261** | -0.0311*** | 0.0344*** |
| | (0.0106) | (0.0098) | (0.0076) |
| $R^2$ | 0.1411 | 0.1115 | 0.0884 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator accounting for journal-year fixed effects (859 groups). Robust standard errors clustered at journal-year level in parentheses. The dependent variables are the shares of an article's geodesic distances of order 1, 2 and 3. $R^2$ is the squared correlation between dependent variable and prediction.
[a] Network variables included in Panel I: number of references in data, percent self-references, average number of references' citations, average number of references' references.
[b] Variables included in Panel II: network variables, paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced.

Table 3: Treatment defined over different online platforms, N=45,553

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| Share online, JSTOR | -0.0165* | -0.0309*** | 0.0279*** |
| | (0.0092) | (0.0085) | (0.0069) |
| Share online, FSO | -0.0543*** | -0.0391*** | 0.0207** |
| | (0.0183) | (0.0148) | (0.0100) |
| $R^2$ | 0.1360 | 0.1041 | 0.0891 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator accounting for journal-year fixed effects (859 groups). Robust standard errors clustered at journal-year level in parentheses. $R^2$ is the squared correlation between dependent variable and prediction. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced, number of references in data, percent self-references, average number of references' citations, average number of references' references.

Table 4: Treatment interacted with time trend, N=45,553

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| Share online | -0.0579*** | -0.0338* | 0.0221 |
| | (0.0222) | (0.0174) | (0.0164) |
| Share online × year | 0.0058 | 0.0005 | 0.0022 |
| | (0.0038) | (0.0030) | (0.0028) |
| $F$-statistic | 4.73 | 5.44 | 10.55 |
| $p$-value | 0.0091 | 0.0045 | 0.0000 |
| $R^2$ | 0.1427 | 0.1122 | 0.0875 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator accounting for journal-year fixed effects (859 groups). Robust standard errors clustered at journal-year level in parentheses. Year in variable "Share online × year" is normalized to zero in 1997. F-statistics and p-values are for joint significance tests on coefficients of "Share online" and "Share online × year". $R^2$ is the squared correlation between dependent variable and prediction. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced, number of references in data, percent self-references, average number of references' citations, average number of references' references.

Table 5: Treatment interacted with journal type, N=45,553

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| Share online × top 5 | -0.0066 | -0.0957** | 0.0473*** |
| | (0.0350) | (0.0374) | (0.0178) |
| Share online × gen. interest | -0.0118 | -0.0460** | 0.0172 |
| | (0.0208) | (0.0207) | (0.0128) |
| Share online × field | -0.0361*** | -0.0136 | 0.0408*** |
| | (0.0126) | (0.0107) | (0.0100) |
| $R^2$ | 0.1387 | 0.1061 | 0.0859 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator accounting for journal-year fixed effects (859 groups). Robust standard errors clustered at journal-year level in parentheses. $R^2$ is the squared correlation between dependent variable and prediction. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced, number of references in data, percent self-references, average number of references' citations, average number of references' references.

Table 6: Citing top 5 journals and own journal, N=45,553

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| I. *Regressions including percent of references to own journal* | | | |
| Share online | -0.0181* | -0.0286*** | 0.0324*** |
| | (0.0106) | (0.0098) | (0.0075) |
| Perc. refs. to own journal | 0.1349*** | 0.0417*** | -0.0335*** |
| | (0.0160) | (0.0126) | (0.0104) |
| $R^2$ | 0.1434 | 0.1131 | 0.0887 |
| II. *Regressions including percent of refs. to "top 5" journals* | | | |
| Share online | -0.0473*** | -0.0294*** | 0.0341*** |
| | (0.0107) | (0.0101) | (0.0077) |
| Percent refs. to top 5 | 0.0817*** | -0.0067 | 0.0013 |
| | (0.0120) | (0.0082) | (0.0079) |
| $R^2$ | 0.1358 | 0.1128 | 0.0884 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator accounting for journal-year fixed effects (859 groups). Robust standard errors clustered at journal-year level in parentheses. $R^2$ is the squared correlation between dependent variable and prediction. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced, number of references in data, percent self-references, average number of references' citations, average number of references' references.

Table 7: Average citation lag, N=45,553

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| Share online | -0.0249** | -0.0262*** | 0.0371*** |
| | (0.0106) | (0.0097) | (0.0076) |
| Average citation lag | -0.0007*** | -0.0028*** | -0.0016*** |
| | (0.0002) | (0.0002) | (0.0002) |
| $R^2$ | 0.1402 | 0.1153 | 0.0940 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator accounting for journal-year fixed effects (859 groups). Robust standard errors clustered at journal-year level in parentheses. $R^2$ is the squared correlation between dependent variable and prediction. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced, number of references in data, percent self-references, average number of references' citations, average number of references' references.

Table 8: Coauthor-group fixed effects, N=21,767

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| Share online | -0.0087 | -0.0189 | 0.0370*** |
| | (0.0179) | (0.0146) | (0.0140) |
| $R^2$ | 0.1817 | 0.1282 | 0.1038 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator. Regressions account for author-group fixed effects (7,307 groups). Standard errors (in parentheses) robust to heteroskedasticity and clustering at coauthor-group level. The dependent variables are the shares of an article's geodesics of order 1, 2 and 3. Further variables included in all regressions: complete set of journal-years indicators, number of references in data, percent self-references, average number of references' citations, average number of references' references, paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced.

Table 9: Alternative identification: Pair-level data, N=406,633

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| I. *Regressions on online indicator and network variables* [a] | | | |
| Online | -0.0817*** | -0.0740** | 0.0389 |
| | (0.0307) | (0.0308) | (0.0242) |
| $R^2$ | 0.1718 | 0.2155 | 0.1345 |
| II. *Regressions on online indicator and further variables* [b] | | | |
| Online | -0.0659** | -0.0833*** | 0.0288 |
| | (0.0294) | (0.0306) | (0.0239) |
| $R^2$ | 0.2371 | 0.2599 | 0.1521 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. The table shows estimates of $\alpha$ from model (4) multiplied by the average number of references, 34.95. All regressions estimated by the OLS within-estimator accounting for cited journal-year fixed effects (1076 groups) and citing journal-year fixed effects (859 groups). Robust standard errors clustered at cited journal-year level in parentheses. The dependent variables are the shares of an article's geodesic distances of order 1, 2 and 3. $R^2$ is the squared correlation between dependent variable and prediction.

[a] Network variables included in Panel I: number of references in data, percent self-references, average number of references' citations, average number of references' references.

[b] Variables included in Panel II: network variables, paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced.

Table 10: Alternative identification: Pair-level data, using $S_{\geq 3}$ as the dependent variable, N=406,633

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Online | 0.1557*** | 0.1492*** | 0.0391 | | 0.1423*** | 0.1490*** |
| | (0.0467) | (0.0453) | (0.1021) | | (0.0451) | (0.0445) |
| Online × year | | | 0.0169 | | | |
| | | | (0.0143) | | | |
| Online × top 5 | | | | 0.3062*** | | |
| | | | | (0.0908) | | |
| Online × gen. interest | | | | 0.2075** | | |
| | | | | (0.0834) | | |
| Online × field | | | | 0.1012* | | |
| | | | | (0.0519) | | |
| $F$-statistic | | | 6.13 | | | |
| $p$-value | | | 0.0022 | | | |
| $R^2$ | 0.1677 | 0.2440 | 0.2441 | 0.2441 | 0.2460 | 0.2494 |

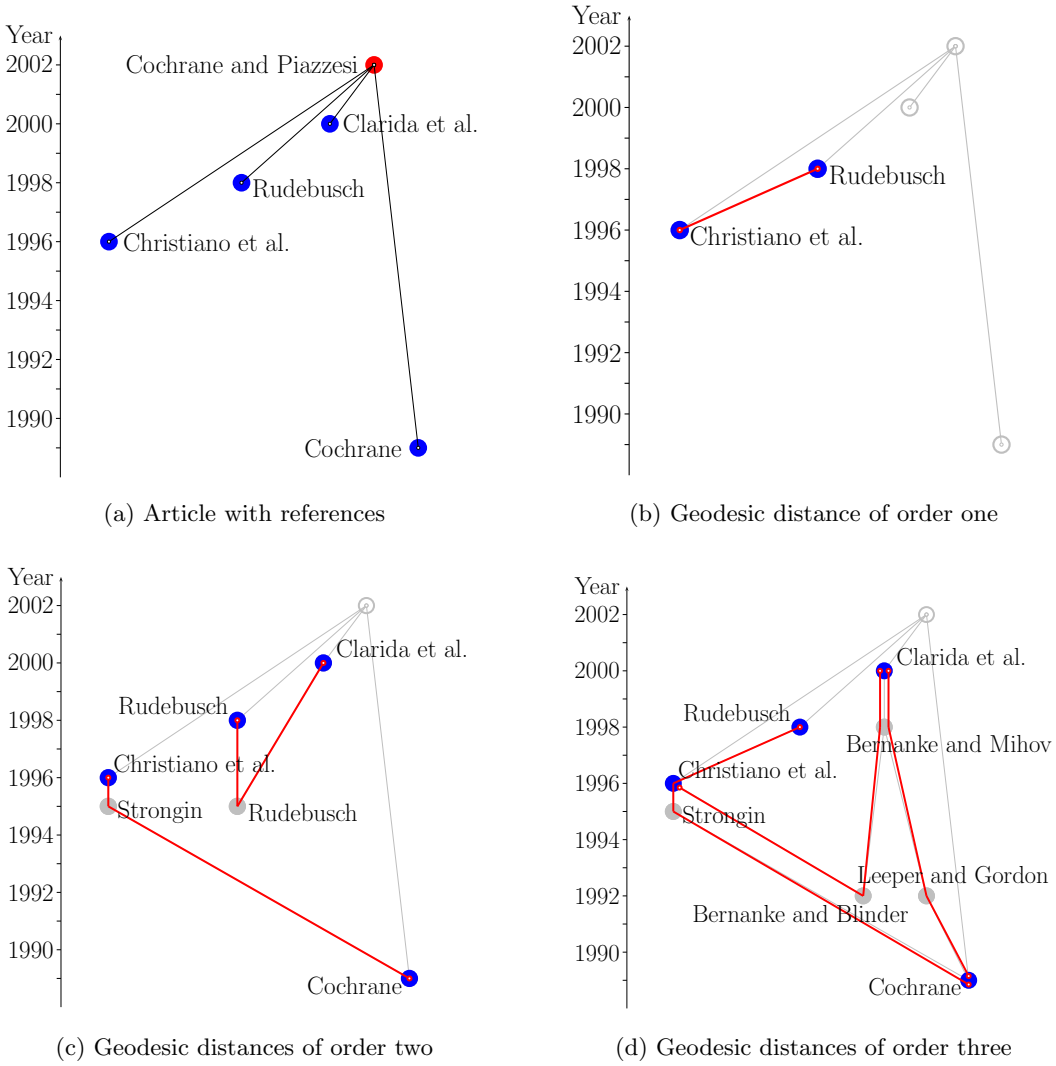
*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. The table shows estimates of $\alpha$ from model (4) multiplied by the average number of references, 34.95. All regressions estimated by the OLS within-estimator accounting for cited journal-year fixed effects (1076 groups) and citing journal-year fixed effects (859 groups). Robust standard errors clustered at cited journal-year level in parentheses. The dependent variable is the shares of a citing article's geodesic distances of order 3 and larger than 3. F-statistic and p-value in column (3) are for the joint test of significance for "Online" and "Online × year". $R^2$ is the squared correlation between dependent variable and prediction.
Column (1): baseline specification with network variables as in Table 2; (2): baseline specification with full set of covariates as in Table 2; (3) specification with time trend interaction as in Table 4; (4): specification with journal type interaction as in Table 5; (5): specification with citing share of top 5 and own journal references as in Table 6; (6): specification with citing article citation lag as in Table 7.

Table 11: Fixed effects regressions of share online on the number of distinct JEL codes, N=45,553

| | 1-digit JEL codes | 2-digit JEL codes | 3-digit JEL codes |
|---|---|---|---|
| I. *Regressions on share online* | | | |
| Share online | 0.2186*** | 0.3309*** | 0.3562*** |
| | (0.0264) | (0.0330) | (0.0377) |
| $R^2$ | 0.0486 | 0.0675 | 0.0933 |
| II. *Regressions on share online and further variables* | | | |
| Share online | 0.1647*** | 0.2565*** | 0.2794*** |
| | (0.0264) | (0.0332) | (0.0377) |
| $R^2$ | 0.0772 | 0.0939 | 0.1138 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator accounting for journal-year fixed effects (859 groups). Robust standard errors clustered at journal-year level in parentheses. The dependent variables are the number of distinct 1-, 2- and 3-digit JEL codes assigned to an article in EconLit. $R^2$ is the squared correlation between dependent variable and prediction. Further variables included in Panel II: number of references in data, percent self-references, average number of references' citations, average number of references' references, paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced.

# Figures




(a) Article with references

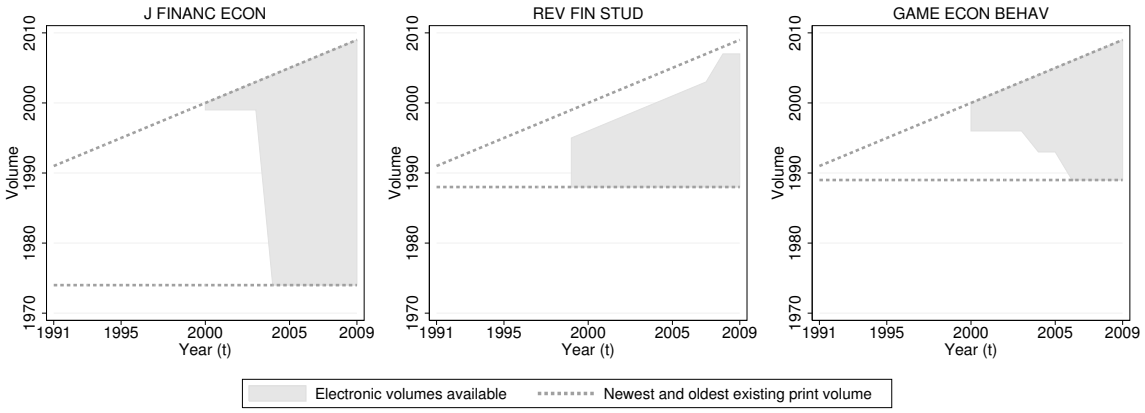
(b) Geodesic distance of order one

(c) Geodesic distances of order two

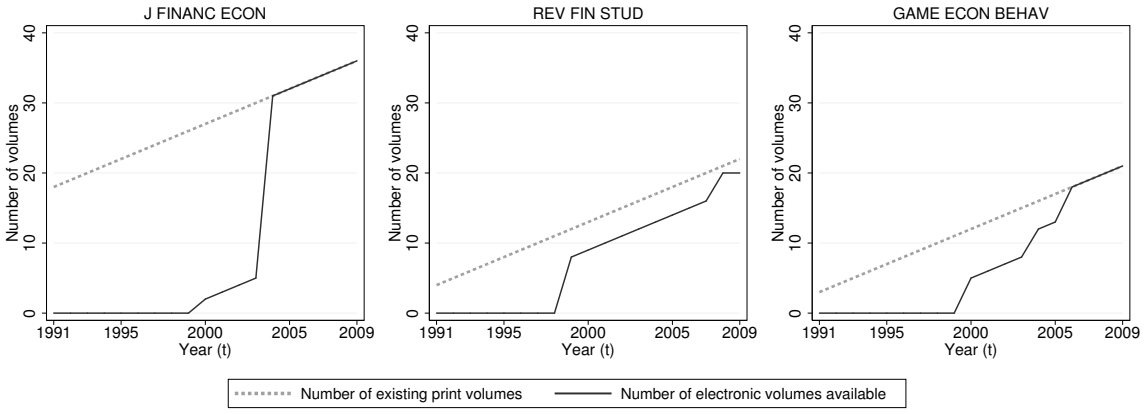
(d) Geodesic distances of order three

Figure 1: **Geodesic distances of an article's references**

**Notes:** The Figure illustrates how geodesic distances of an article's references are obtained, using the article by John Cochrane and Monica Piazzesi, "The Fed and Interest Rates: A High-Frequency Identification", *American Economic Review, Papers and Proceedings*, May 2002, Volume 92, Issue 2, pp. 90-95. Panel (a) plots the article as a red node and the four references identified in the data as blue nodes. Panels (b), (c) and (d) plot shortest back-in-time citation paths between blue nodes (geodesic distances) as red lines. Blue nodes' references relevant for these paths are plotted in grey.



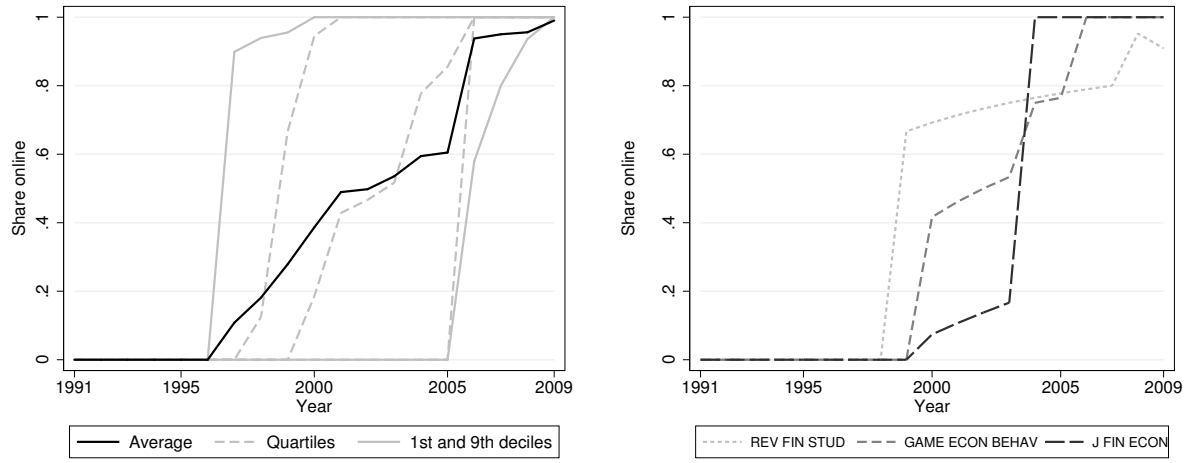

(a) Available print and electronic volumes




(b) Number of available print and electronic volumes

Figure 2: **Print and electronic volumes of three selected journals over time**

**Notes:** The Figure illustrates the availability of print volumes and electronic volumes for three journals –Journal of Financial Economics, Review of Financial Studies, and Games and Economic Behavior– over the period 1991-2009. In panel (a), the bottom dotted lines represent the oldest print volumes available (corresponding to the year when the journals were founded) and the top dotted lines represent the newest existing print volume (corresponding to the current year in the x-axis). The area colored in gray represents volumes accessible in full text on the internet. Panel (b) depicts the number of available print volumes (dotted lines) and electronic volumes (solid lines).




(a) Share online of all journals

(b) Share online of three selected journals

Figure 3: **Share online of journals**

**Notes:** The Figure illustrates "Share online" for various journals. Share online is the share of volumes available electronically in all volumes (available in print). Panel (a) plots selected descriptive statistics of share online of all journals. Panel (b) plots share online for the three selected journals from Figure 2.

# Appendix

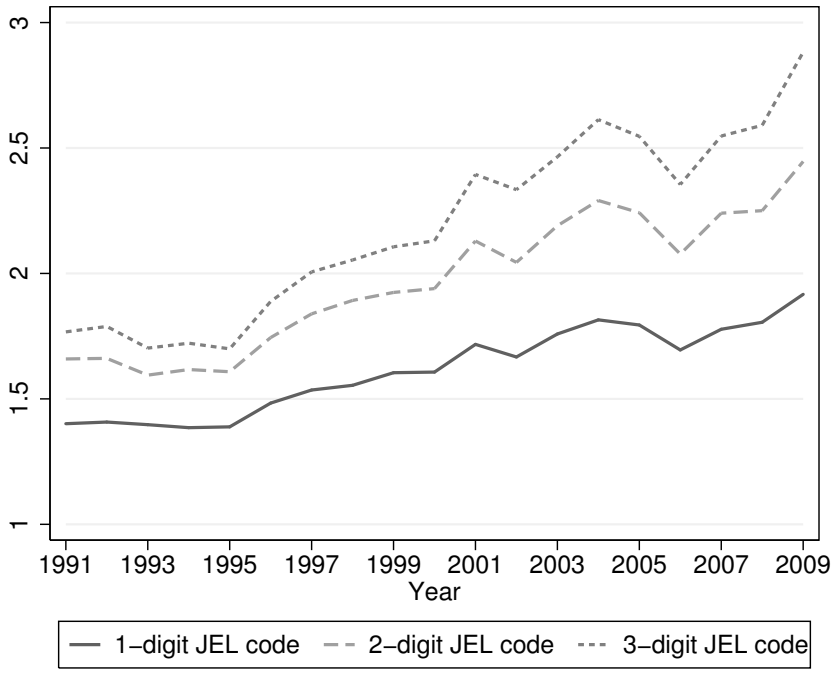



Figure A.1: **Average number of distinct JEL codes per article over time**

**Notes:** The figure plots the average number of one, two and three digit JEL codes. The sample includes the 45,553 articles published between 1991-2009 in the considered journals.

Table A.1: Means and standard deviations (SD) of selected variables

| Variable | Mean | SD, overall | SD, between | SD, within |
|---|---|---|---|---|
| $S_1$ | 0.2464 | 0.2055 | 0.0606 | 0.1963 |
| $S_2$ | 0.2707 | 0.1792 | 0.0643 | 0.1697 |
| $S_3$ | 0.1983 | 0.1523 | 0.0455 | 0.1466 |
| $S_{>3}$ | 0.2845 | 0.2511 | 0.1015 | 0.2340 |
| 1-digit JEL code | 1.6436 | 0.7526 | 0.3504 | 0.6858 |
| 2-digit JEL code | 2.0104 | 0.9958 | 0.4963 | 0.8916 |
| 3-digit JEL code | 2.2474 | 1.1404 | 0.6173 | 0.9940 |
| Share online | 0.5188 | 0.4188 | 0.4049 | 0.1258 |
| Number of observations (articles) | | | 45,553 | |
| Number of groups (journal-years) | | | 859 | |
| Average number of observations per group | | | 50.03 | |

*Notes:* "$S_l$" denotes the share of geodesics between an article's references of order $l$. "$l$-digit JEL code" denotes the number of distinct $l$-digit JEL codes assigned to an article by EconLit. "Share online" is the share of an article's relevant literature that was accessible online as defined in (2). "SD, between" and "SD, within" are standard deviations calculated between and within journal-year groups.

Table A.2: Estimates of models for the probability of being among the 5% most cited articles (logit regressions with journal-years fixed effects)

| Among top 5%... | ...after 20 yrs. | ...after 30 yrs. | ...after 40 yrs. |
|---|---|---|---|
| I. *Regressions controlling for network variables* | | | |
| $S_{\geq 3}$ | 0.5939*** | 0.6012*** | 1.4479*** |
| | (0.1111) | (0.1693) | (0.5125) |
| N | 12,854 | 4,591 | 828 |
| $R^2$ | 0.3453 | 0.3503 | 0.4377 |
| II. *Regressions controlling for network and further variables* | | | |
| $S_{\geq 3}$ | 0.2875** | 0.4079** | 1.3928** |
| | (0.1197) | (0.1876) | (0.5480) |
| N | 12,854 | 4,591 | 828 |
| $R^2$ | 0.3733 | 0.3673 | 0.4625 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by fixed effects logit conditional likelihood estimation accounting for journal-year fixed effects. Results show logit coefficients. Robust standard errors clustered at journal-year level in parentheses. $R^2$ is McFadden's pseudo R-squared. The dependent variables are indicators equal to one if the article is in the 95th percentile of citations 20, 30 and 40 years after publication among articles published in the same year. The key explanatory variable, $S_{\geq 3}$, is the share of geodesics of order greater than two among geodesics between an article's references. Network control variables: number of references in data, percent self-references, average number of references' citations, average number of references' references. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced.

Table A.3: Baseline regression (Panel II of Table 2) – Full output, N=45,553

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| Share online | -0.0261** | -0.0311*** | 0.0344*** |
| | (0.0106) | (0.0098) | (0.0076) |
| Proceedings paper | -0.0052 | -0.0034 | 0.0070** |
| | (0.0039) | (0.0033) | (0.0029) |
| No. authors | -0.0054*** | 0.0002 | 0.0024*** |
| | (0.0012) | (0.0009) | (0.0009) |
| No. pages | -0.0016*** | 0.0002 | 0.0009*** |
| | (0.0001) | (0.0001) | (0.0001) |
| No. references | -0.0023*** | -0.0026*** | -0.0003*** |
| | (0.0001) | (0.0001) | (0.0001) |
| No. journals referenced | -0.0191*** | -0.0027*** | 0.0075*** |
| | (0.0006) | (0.0006) | (0.0004) |
| No. refs. in data | 0.0031*** | 0.0085*** | 0.0015*** |
| | (0.0002) | (0.0002) | (0.0002) |
| Perc. self-refs. | 0.1221*** | -0.0053 | -0.0380*** |
| | (0.0108) | (0.0075) | (0.0065) |
| Avg. ref.'s refs. $\times 10^{-1}$ | 0.0094*** | 0.0205*** | 0.0096*** |
| | (0.0013) | (0.0013) | (0.0010) |
| Avg. ref.'s cit. $\times 10^{-2}$ | 0.0037*** | 0.0026*** | -0.0008* |
| | (0.0006) | (0.0005) | (0.0004) |
| $R^2$ | 0.1411 | 0.1115 | 0.0884 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator accounting for journal-year fixed effects (859 groups). Robust standard errors clustered at journal-year level in parentheses. $R^2$ is the squared correlation between dependent variable and prediction.

Table A.4: Different lags of treatment, N=45,553

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| I. *Regressions using contemporaneous treatment* | | | |
| Share online | -0.0399*** | -0.0244** | 0.0308*** |
| | (0.0112) | (0.0103) | (0.0080) |
| $R^2$ | 0.1374 | 0.1140 | 0.0880 |
| II. *Regressions using treatment lagged by one year* | | | |
| Share online | -0.0345*** | -0.0270*** | 0.0357*** |
| | (0.0115) | (0.0100) | (0.0076) |
| $R^2$ | 0.1385 | 0.1128 | 0.0881 |
| III. *Regressions using treatment lagged by two years* | | | |
| Share online | -0.0261** | -0.0311*** | 0.0344*** |
| | (0.0106) | (0.0098) | (0.0076) |
| $R^2$ | 0.1411 | 0.1115 | 0.0884 |
| IV. *Regressions using treatment lagged by three years* | | | |
| Share online | -0.0130 | -0.0317*** | 0.0337*** |
| | (0.0093) | (0.0080) | (0.0068) |
| $R^2$ | 0.1440 | 0.1123 | 0.0886 |
| V. *Regressions using treatment lagged by four years* | | | |
| Share online | -0.0000 | -0.0344*** | 0.0198*** |
| | (0.0086) | (0.0075) | (0.0061) |
| $R^2$ | 0.1455 | 0.1126 | 0.0883 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator accounting for journal-year fixed effects (859 groups). Robust standard errors clustered at journal-year level in parentheses. $R^2$ is the squared correlation between dependent variable and prediction. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced, number of references in data, percent self-references, average number of references' citations, average number of references' references.

Table A.5: Alternative treatments, N=45,553

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| I. *Regressions using average share online* | | | |
| Avg. share online | -0.0084 | -0.0433*** | 0.0259*** |
| | (0.0097) | (0.0088) | (0.0072) |
| $R^2$ | 0.1447 | 0.1076 | 0.0891 |
| II. *Regressions using share of references online* | | | |
| Share refs. online | 0.0096 | -0.0266*** | 0.0041 |
| | (0.0087) | (0.0073) | (0.0055) |
| $R^2$ | 0.1460 | 0.1147 | 0.0856 |
| III. *Regressions using number of volumes online* | | | |
| Vols. online | -0.0001 | -0.0389*** | 0.0087** |
| | (0.0051) | (0.0045) | (0.0036) |
| $R^2$ | 0.1455 | 0.1156 | 0.0865 |
| IV. *Regressions using average no. of volumes online* | | | |
| Avg. vols. online | -0.0080 | -0.0316*** | 0.0165*** |
| | (0.0057) | (0.0050) | (0.0043) |
| $R^2$ | 0.1450 | 0.1143 | 0.0888 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. Depicted effects are for a change from print literature to completely available online literature; for panels III and IV, this was evaluated at the mean number of references (27) in the year 2009. All regressions estimated by the OLS within-estimator accounting for journal-year fixed effects (859 groups). Robust standard errors clustered at journal-year level in parentheses. $R^2$ is the squared correlation between dependent variable and prediction. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced, number of references in data, percent self-references, average number of references' citations, average number of references' references.

Table A.6: Alternative online access data, N=33,015 (source: McCabe and Snyder, 2014 [MS])

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| I. *MS online access data* | | | |
| Share online | -0.0273** | -0.0368*** | 0.0315*** |
| | (0.0116) | (0.0103) | (0.0087) |
| $R^2$ | 0.1357 | 0.1044 | 0.0776 |
| II. *MS online access data & partial access variable* | | | |
| Share online | -0.0397*** | -0.0385*** | 0.0341*** |
| | (0.0124) | (0.0120) | (0.0093) |
| $R^2$ | 0.1314 | 0.1032 | 0.0774 |
| III. *Default online access data, same estimation sample* | | | |
| Share online | -0.0361*** | -0.0381*** | 0.0322*** |
| | (0.0106) | (0.0111) | (0.0080) |
| $R^2$ | 0.1336 | 0.1038 | 0.0773 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator accounting for journal-year fixed effects (673 groups). Robust standard errors clustered at journal-year level in parentheses. $R^2$ is the squared correlation between dependent variable and prediction. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced, number of references in data, percent self-references, average number of references' citations, average number of references' references. Additionally, Panel II includes a control variable indicating the share of volumes with partial online accessibility.

Table A.7: Refining journal fixed effects

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| I. *Journal-document-type-year fixed effects, N=45,553* | | | |
| Share online | -0.0253** | -0.0299*** | 0.0349*** |
| | (0.0106) | (0.0099) | (0.0076) |
| $R^2$ | 0.1411 | 0.1118 | 0.0879 |
| II. *Journal-issue-year fixed effects, N=44,937* | | | |
| Share online | -0.0187* | -0.0296*** | 0.0325*** |
| | (0.0109) | (0.0103) | (0.0079) |
| $R^2$ | 0.1410 | 0.1123 | 0.0857 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator. Panel I accounts for journal-document-type-year fixed effects (1,456 groups), panel II for journal-issues-years fixed effects (4,817 groups). Robust standard errors clustered at journal-year level in parentheses. $R^2$ is the squared correlation between dependent variable and prediction. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced, number of references in data, percent self-references, average number of references' citations, average number of references' references.

Table A.8: References' age distribution, N=45,553

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| I. *Average citation lag and cit. lag's standard deviation* | | | |
| Share online | -0.0252** | -0.0255*** | 0.0372*** |
| | (0.0107) | (0.0098) | (0.0076) |
| Average citation lag | -0.0004 | -0.0037*** | -0.0017*** |
| | (0.0003) | (0.0003) | (0.0002) |
| Lag's std. dev.$\times 10^{-2}$ | -0.0010 | 0.0027*** | 0.0005 |
| | (0.0006) | (0.0004) | (0.0004) |
| $R^2$ | 0.1406 | 0.1151 | 0.0940 |
| II. *Median citation lag* | | | |
| Share online | -0.0258** | -0.0271*** | 0.0366*** |
| | (0.0106) | (0.0098) | (0.0076) |
| Median citation lag | -0.0002 | -0.0026*** | -0.0015*** |
| | (0.0002) | (0.0002) | (0.0002) |
| $R^2$ | 0.1408 | 0.1138 | 0.0925 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by the OLS within-estimator accounting for journal-year fixed effects (859 groups). Robust standard errors clustered at journal-year level in parentheses. $R^2$ is the squared correlation between dependent variable and prediction. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced, number of references in data, percent self-references, average number of references' citations, average number of references' references.

Table A.9: Alternative identification: Pair-level data, N=522,692

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| I. *Regressions on online indicator and network variables* [a] | | | |
| Online | -0.0394 | -0.0397 | 0.0282 |
| | (0.0271) | (0.0272) | (0.0215) |
| $R^2$ | 0.1665 | 0.2139 | 0.1411 |
| II. *Regressions on online indicator and further variables* [b] | | | |
| Online | -0.0271 | -0.0463* | 0.0202 |
| | (0.0259) | (0.0269) | (0.0212) |
| $R^2$ | 0.2306 | 0.2578 | 0.1586 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. The table shows estimates of $\alpha$ from model (4) multiplied by the average number of references, 34.37. All regressions estimated by the OLS within-estimator accounting for cited journal-year fixed effects (1592 groups) and citing journal-year fixed effects (859 groups). Robust standard errors clustered at cited journal-year level in parentheses. The dependent variables are the shares of an article's geodesic distances of order 1, 2 and 3. $R^2$ is the squared correlation between dependent variable and prediction.

[a] Network variables included in Panel I: number of references in data, percent self-references, average number of references' citations, average number of references' references.

[b] Variables included in Panel II: network variables, paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced.

Table A.10: Poisson fixed effects regressions of percent online on the number of distinct JEL codes, N=45,553

| | 1-digit | 2-digits | 3-digits |
|---|---|---|---|
| I. *Regressions on percent online* | | | |
| Share online | 0.1345*** | 0.1687*** | 0.1622*** |
| | (0.0168) | (0.0173) | (0.0177) |
| $R^2$ | 0.0644 | 0.0721 | 0.0812 |
| II. *Regressions on percent online and further variables* | | | |
| Share online | 0.1042*** | 0.1356*** | 0.1313*** |
| | (0.0170) | (0.0178) | (0.0180) |
| $R^2$ | 0.0660 | 0.0742 | 0.0836 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by fixed effects Poisson pseudo-likelihood estimation accounting for journal-year fixed effects (859 groups). Robust standard errors clustered at journal-year level in parentheses. The dependent variables are the number of distinct 1-, 2- and 3-digit JEL codes assigned to an article in EconLit. $R^2$ is McFadden's pseudo-R-squared. Further variables included in Panel II: number of references in data, percent self-references, average number of references' citations, average number of references' references, paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced.

Table A.11: Alternative identification for JEL code regressions: Pair-level data, N=406,633

| | 1-digit | 2-digits | 3-digits |
|---|---|---|---|
| I. *Regressions on share online,*[a] | | | |
| Online | 0.5903*** | 0.7723*** | 0.8601*** |
| | (0.1802) | (0.2204) | (0.2496) |
| $R^2$ | 0.1691 | 0.1979 | 0.2356 |
| II. *Regressions on share online and further variables*[b] | | | |
| Percent references online | 0.5232*** | 0.6856*** | 0.7504*** |
| | (0.1770) | (0.2178) | (0.2465) |
| $R^2$ | 0.1850 | 0.2107 | 0.2486 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. The table shows estimates of $\alpha$ from model (4) multiplied by the average number of references, 34.95. All regressions estimated by the OLS within-estimator accounting for cited journal-year fixed effects (1076 groups) and citing journal-year fixed effects (859 groups). Robust standard errors clustered at cited journal-year level in parentheses. The dependent variables are the number of distinct 1-, 2- and 3-digit JEL codes assigned to an article in EconLit. $R^2$ is the squared correlation between dependent variable and prediction. Further variables included in Panel II: number of references in data, percent self-references, average number of references' citations, average number of references' references, paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced.

Table A.12: Estimates of models for an article's total number of citations (OLS regressions with journal-years fixed effects)

| | Cit. after 20 yrs. | Cit. after 30 yrs. | Cit. after 40 yrs. |
|---|---|---|---|
| I. *Regressions controlling for network variables* | | | |
| $S_{\geq 3}$ | 6.6570*** | 12.6114*** | 22.5138*** |
| | (1.6990) | (4.4322) | (5.7239) |
| N | 21,501 | 7,718 | 1,539 |
| $R^2$ | 0.0235 | 0.0140 | 0.0302 |
| II. *Regressions controlling for network and further variables* | | | |
| $S_{\geq 3}$ | 1.4051 | 7.8459** | 18.7593*** |
| | (1.5469) | (3.9005) | (5.9326) |
| N | 21,501 | 7,718 | 1,539 |
| $R^2$ | 0.0432 | 0.0242 | 0.0470 |

*Notes:* *, ** and *** indicate statistical significance at 10%, 5% and 1% level. All regressions estimated by fixed effects OLS estimation accounting for journal-year fixed effects. Robust standard errors clustered at journal-year level in parentheses. $R^2$ is the squared correlation between dependent variable and prediction. The dependent variables are the total number of citations received by an article 20, 30 and 40 years after publication. The key explanatory variable is the share of geodesic distances of order greater than two among all pairwise geodesic distances of an article's references. Network control variables: number of references in data, percent self-references, average number of references' citations, average number of references' references. Further control variables: paper-and-proceedings indicator, number of authors, number of pages, number of references, number of journals referenced.

Table A.13: Descriptive statistics of data used in estimation

| Variable | Mean | Std. Dev. | Min. | Max. |
|---|---|---|---|---|
| $S_1$ | 0.2464 | 0.2055 | 0 | 1 |
| $S_2$ | 0.2707 | 0.1792 | 0 | 1 |
| $S_3$ | 0.1983 | 0.1523 | 0 | 1 |
| $S_{>3}$ | 0.2845 | 0.2511 | 0 | 1 |
| 1-digit JEL code | 1.6436 | 0.7526 | 1 | 6 |
| 2-digit JEL code | 2.0104 | 0.9958 | 1 | 7 |
| 3-digit JEL code | 2.2474 | 1.1404 | 1 | 8 |
| Share online | 0.5188 | 0.4188 | 0 | 1 |
| Share online, JSTOR | 0.4538 | 0.3721 | 0 | 0.9894 |
| Share online, FSO | 0.207 | 0.2742 | 0 | 1 |
| Share online, no lag | 0.624 | 0.4068 | 0 | 1 |
| Share online, 1-year lag | 0.5731 | 0.4169 | 0 | 1 |
| Share online, 3-year lag | 0.4556 | 0.4086 | 0 | 1 |
| Average percent online | 0.4438 | 0.3872 | 0 | 1 |
| No. volumes online | 189.6986 | 200.1333 | 0 | 1055 |
| Average no. volumes online | 30.765 | 27.7792 | 0 | 122 |
| Percent references online | 0.4382 | 0.4009 | 0 | 1 |
| Document type: article | 0.9063 | 0.2914 | 0 | 1 |
| Doc. type: proceedings paper | 0.0937 | 0.2914 | 0 | 1 |
| No. authors | 1.8287 | 0.8609 | 1 | 26 |
| No. pages | 20.1024 | 10.1193 | 1 | 96 |
| No. journals referenced | 5.6985 | 2.8131 | 1 | 21 |
| No. references | 27.1517 | 16.2991 | 2 | 537 |
| No. refs. in data | 11.5066 | 8.1799 | 2 | 222 |
| Percent self-references | 0.0859 | 0.1431 | 0 | 1 |
| Top 5 journal | 0.1254 | 0.3312 | 0 | 1 |
| General interest journal | 0.2603 | 0.4388 | 0 | 1 |
| Field journal | 0.6143 | 0.4868 | 0 | 1 |
| Percent refs. to top 5 | 0.1881 | 0.1396 | 0 | 1 |
| Perc. refs. to own journal | 0.071 | 0.093 | 0 | 1 |
| Average citation lag | 11.6325 | 5.737 | 0 | 100.2258 |
| Median citation lag | 8.8366 | 5.1092 | 0 | 89 |
| Std. dev. of cit. lag | 126.3215 | 276.2136 | 0 | 9528.876 |
| Average ref.'s references | 27.8549 | 11.3983 | 0.5 | 290.5 |
| Average ref.'s citations | 227.5147 | 241.517 | 1 | 4104 |
| N | | 45,553 | | |

Table A.14: Alphabetic list of the selected journals

| No. | Journal | No. of items |
|---|---|---|
| 1 | American Economic Review | 11,246 |
| 2 | Bell Journal of Economics | 542 |
| 3 | Econometric Theory | 1,288 |
| 4 | Econometrica | 6,039 |
| 5 | Economic Inquiry | 1,892 |
| 6 | Economic Journal | 9,397 |
| 7 | Economic Theory | 1,402 |
| 8 | Economics Letters | 7,115 |
| 9 | European Economic Review | 2,992 |
| 10 | Games and Economic Behavior | 1,436 |
| 11 | International Economic Review | 1,898 |
| 12 | International Journal of Game Theory | 745 |
| 13 | Journal of Applied Econometrics | 1,029 |
| 14 | Journal of Business Economic Statistics | 1,334 |
| 15 | Journal of Development Economics | 2,356 |
| 16 | Journal of Econometrics | 2,705 |
| 17 | Journal of Economic Behavior and Organization | 2,464 |
| 18 | Journal of Economic Dynamics and Control | 2,072 |
| 19 | Journal of Economic Growth | 146 |
| 20 | Journal of Economic History | 10,355 |
| 21 | Journal of Economic Literature | 6,916 |
| 22 | Journal of Economic Perspectives | 1,329 |
| 23 | Journal of Economic Theory | 3,359 |
| 24 | Journal of Environmental Economics and Management | 1,434 |
| 25 | Journal of Finance | 6,979 |
| 26 | Journal of Financial and Quantitative Analysis | 1,988 |
| 27 | Journal of Financial Economics | 1,629 |
| 28 | Journal of Financial Intermediation | 290 |
| 29 | Journal of Health Economics | 1,208 |
| 30 | Journal of Human Resources | 1,786 |
| 31 | Journal of International Economics | 2,305 |
| 32 | Journal of Labor Economics | 834 |
| 33 | Journal of Mathematical Economics | 1,228 |
| 34 | Journal of Monetary Economics | 2,056 |
| 35 | Journal of Political Economy | 4,880 |
| 36 | Journal of Public Economics | 2,633 |
| 37 | Journal of Risk and Uncertainty | 566 |
| 38 | Journal of the European Economic Association | 331 |
| 39 | Journal of Urban Economics | 1,684 |
| 40 | Oxford Bulletin of Economics and Statistics | 1,449 |
| 41 | Quarterly Journal of Economics | 2,677 |
| 42 | RAND Journal of Economics | 1,113 |
| 43 | Review of Economic Studies | 2,338 |
| 44 | Review of Economics and Statistics | 4,429 |
| 45 | Review of Financial Studies | 916 |
| 46 | Scandinavian Journal of Economics | 1,547 |
| 47 | Social Choice and Welfare | 1,143 |
| 48 | Swedish Journal of Economics | 326 |
| 49 | Western Economic Journal | 672 |
| 50 | World Bank Economic Review | 647 |
| | Total number of items 1955-2009 | 129,145 |

*Notes*: "Bell Journal of Economics" includes its predecessor "The Bell Journal of Economics and Management Science" 1970-1974. "Oxford Bulletin of Economics and Statistics" includes its predecessor "Oxford University Bulletin of the Institute of Economics and Statistics" 1939-1972. Our sample includes three historical, non-successive journals: "Bell Journal of Economics" 1970-1983, "Swedish Journal of Economics" 1965-1975 and the "Western Economic Journal" 1962-1972. We ignore the non-English predecessor of the "Swedish Journal of Economics" - "Ekonomisk Tidskrift" - which goes back to 1899.
The list includes all journals considered in the standard Tilbourg ranking as well as the list considered in Palacios-Huerta and Volij (2004).
Some isolated publication years are missing, since they are not included in the Web of Science.