Echo Chambers in Science?

Lanu Kim, Jevin West, and Katherine Stovel

January 2016

N:B: This early draft presents the first findings from a new project that looks at the potential effects of technology and academic search engines on the focus and breadth of academic exploration, as reflected in citation patterns. Inspired by Geertz' distinction between intensive and extensive search, we investigate whether citation patterns in two disciplines have become more concentrated or distributed in the wake of the widespread use of new search technologies. The text of this draft is rough (and in places incomplete!!) since we have just completed our first set of analyses, but we are quite enthusiastic about the prospect of significant insight coming out of this project in the coming months.

Abstract

This paper examines whether digitization and the rise of integrated academic search engines have transformed how researchers engage with previous literature, a critical component of modern scientific practice. Among technological advancements, we particularly focus on the recent emergence of academic search engines such as Google Scholar, because these services are provided based on proprietary algorithms that actively interfere in authors' search process. While the impact of general recommender systems has been widely noted, the effect of academic recommender systems on scientific practice has not been fully examined. Using the comprehensive Web of Science database covering a wide range of publications and the citation links between them, we focus on yearly changes in the the citing behavior of researchers in two relatively similar disciplines, Sociology and Social Work, between 1998 to 2014. We document three temporal changes in researchers' behavior. First, researchers' citations in both disciplines have become more expansive since 2005 and stable after 2010. Second, controlling for a measure of journal prestige, the impact of a paper-based popularity measure, the cumulative previous citation count, has increased in both Sociology and Social Work. Third, more papers published in lower-tier journals are now cited than prior to 2005, and the variability of citation counts among papers published in the same journal has also increased. Based on three findings, we see some evidence that the digitization of science has democratized the exposure of prior research and weakened journals' role as gatekeepers. Nevertheless, the increasing importance of prior citations suggests a competing trend is also occurring that may create an echo chamber centered on small numbers of highly cited papers.

1 Introduction

This article examines the impact of digitization and the proliferation of academic search engines on the ways in which scholars engage with the literature. Accessing prior scientific knowledge is a critical component of modern scientific practice, and technological developments over the past several decades have revolutionized how scientists discover and cite past research [9]. Among these technological developments, the recent emergence of academic search engine ,such as Google Scholar and Microsoft Academic Search, are of particular interest because these services are powered by black-box, proprietary algorithms that attempt to actively anticipate users' needs rather than simply listing papers indexed by keyword. The question motivating our research is, are these tools enhancing scholars access to a wide range of potentially relevant prior work, or are they concentrating scholars' gaze onto an ever smaller set of "star" papers? Focusing on temporal changes in patterns of citation, we investigate whether the new technologies are associated with an expansion or contraction in the corpus of prior research.

Most scientists now access their literature through online search engines and digital libraries. Rare is the scientist who walks to the library and peruses the journal shelves for new papers. Most researchers now rely on Google Scholar, PubMed, Web of Science, and various other search engines for finding and navigating the literature[1]. Collectively, are we reading more or less of the literature in this new digital environment? Are the citation rich becoming more rich? Or are papers published in lower tier journals receiving more citations than they were prior to the digital divide? What impact is Google Scholar and recommender systems, in general, having on what papers we find, read and cite? Ultimately, we will investigate these questions at the individual and population level. For this paper, we focus on the population-level behavior via citations before, during and after these digital transitions.

When using search engines, users tend to click on the top results and rarely move to subsequent pages. A recent study[2] showed that users click on the top two positions more than 50% of the time when using Google. This search behavior is likely similar when using Google Scholar. Also, Lerman ([16]) found that the presentation order of search results greatly affect how users allocate attention due to human cognitive biases. Thus, if there exists algorithmic bias (or even just bugs in the retrieval code), this could have adverse affects on the search capacity, especially given the increased reliance on Google Scholar as the primary way of accessing the literature. Depending on how the algorithms order

[1]There is a distinction between searching for a paper when one knows the title, author or DOI (Digatal Oject Identifier) and when one is "navigating" the literature with no one paper in mind. The latter is the kind of search that motivates our research question. The nearly instantaneous access to millions of papers is without question a good thing for science. What is less clear is whether the current search engines, library portals and recommender systems are good for science. We want to investigate the role these new technologies and recommender systems are having on what is being read and subsequently cited.

[2]https://searchenginewatch.com/sew/study/2276184/no-1-position-in-google-gets-33-of-search-traffic-study

results, some papers will receive less than their 'fair' share of hits and others will receive more than their 'fair' share. Some perspectives and relevant findings will go unnoticed, while 'top ten' lists will echo louder the same highly cited papers. In a classic 'Matthew Effect,' scholars will cite the same papers that show up in the top ten lists and those papers will subsequently receive more citations and then be weighted higher in the search results.

If researchers are finding (and subsequently citing) the same papers from the same search engines and the same recommender systems, this could have adverse effects on what is read in the literature. Recent studies have pointed to the potential myopia of science [19]. Second and third tier journals may be accessed less and less with sleeping beauties never waking up [14] (although the negative effects could be balanced with search engines uncovering the less cited papers). In addition, people's careers depend on whether academic engines show their papers on the first page rather than the 10th page of search results. Those in the first couple pages have a much greater chance of being cited, which could lead to promotions.

Though the impact of academic recommender systems on the consumption of scientific research has not been examined directly, the influence of the recommender system embedded in large online commercial websites such as Amazon, Target, Spotify, and Netflix on consumption patterns have been studied extensively. The results of these studies reveal complex and contradictory effects: on the one hand, studies of online clothing markets [5] and the video rental market [26] have found that consumers' use of recommendation engines increases the share of niche product consumption, while other studies suggest that online markets driven by recommender systems lead to convergence around a smaller number of popular items [20, 8, 10].

Despite the significance of studying the rise of academic recommender systems, there are methodological issues associated with using bibliographic analysis to investigate the consequences of search engines on scientific practice. The key problem is that we do not observe scholars behavior directly, and thus do not know how, exactly, scientists locate the research cited in their bibliography – or whether this has changed in response to new technologies. Our strategy in this article is to attempt to isolate the effect of technology from other confounding variables – such as the citation norm of fields or increased number of publications – by comparing temporal changes in the relative influence of journal and paper. One of the characteristics of the new systems is that they provide paper-specific information such as the number of received citations and hyperlinks from and toward a paper. Therefore, if the impact of journal prestige on citation has declined while the impact of prior citations to a given paper increased after the academic recommender system has been popularized, we can argue that this new system is changing the behavior of researchers. A secondary problem is that there is a great deal of variation in citation norms by discipline [12]. We control disciplinary norms by conducting our analyses separately for different disciplines, and investigate whether the same behavior pattern is observed across fields.

In this paper, we also compare the relative influence of journals and papers.

This allows us to rethink the role of journals in the Google Scholar era. Prior to the advent of digitization, journals played a central role in the scientific process, both evaluating research (through peer-review and the editorial process) and serving as an efficient filter for the search process. While the gatekeeping role of journals has been well documented, journals' impact on the search process is less well appreciated. However, we believe that both the physical location of journal archives on library shelves and personal subscriptions facilitated individual scholars being familiar with the research publish in a particular set of journals – all while increasing the difficulty of learning about, let alone gaining access to, scholarship published in other outlets. Digitization made the process of accessing a known paper far easier, but it initially did littler to improve search, and the curatorial role of journals remained. A consequence of the pre-academic recommender systems era is that papers published in high profile or well distributed journals were likely to be seen, and hence cited, more than papers published in journals with smaller subscription bases or lower reputations. In the Google Scholar era, however, papers' positions in various electronic archives (and the algorithms used to access these archives) may be increasingly decoupled from the journal they are published in, while paper specific features (such as prior citation, or authorship) may play an increasing role in their visibility to future scholars. Thus, comparing changes in journal-level and paper-level measures enables us to separate two effects of technology, digitization and the rise of academic recommender system.

2 Background/Literature review

Google Scholar was launched in November 2004 with a goal to create one efficient search engine where scholars as well as general public can find scholarly information from all disciplines and languages [11]. Google Scholar begins with searching keywords as well as showing citation counts in the search results. In a relatively short amount of time, it has become one of the most important search tools for researchers' access to the literature. It has continued to develop features, including author-level metrics and links to Web of Science citations. Google Scholar has also created its own index for 'popular' papers and recently added an application to automatically search articles while browsing the web. How Google Scholar ranks papers, however, has not been disclosed. This is a black box, but researchers have tried to reverse engineer the results beel2009google. Based on these studies, there is evidence that it highly weights citation counts, which means that highly cited papers are more likely to be shown in top positions. Beel and Gipp [3] concluded that Google Scholar is more suitable "when searching for standard literature rather than gems, the latest trends, or article by authors advancing a different view from the mainstream", and more susceptible to increasing the Matthew effect. Nevertheless, the consensus about how to construct search algorithms and how to evaluate the performance of them (e.g., [25]) has not been reached yet [2]. Also, search outcomes are frequently changing in response to algorithmic updates or A-B

testing.

As search engines become popularized, the necessity of a paper-level metric instead of journal-based influence factor has increased as well. Limitations of journal-level metrics are well described in Lariviere et al.'s recent research [15]; they found that even though journals have a different spectrum of journal influences, the citation distributions of published papers largely overlap each other indicating that the quality of a paper cannot be inferred from journal's status. In line with this argument, Altmetric is an example of a new paper-level metric in the Web 2.0 era that counts things like tweets from social media [22].

Although there has been a lively discussion about how to construct effective and stable search algorithms and how to measure the significance of academic outcome in the Web 2.0 era, assessing the impact of technological development on researchers' behavior has been relatively neglected. Some have proposed explanations about temporal changes in citation distributions that could be driven by Google Scholar ([23, 21]), but the limitation is that the studies only include classic papers and does not appropriately control for the numerous confounding variables that could be influencing citing behavior. In spite of these limitations, these two analyses suggest a possibility that the development of new technology might bright about more concentrated citation distributions. It is not only recommender systems, but the digitization of journals might also contribute to the convergence of citations by accelerating consensus through quick communication [9]. In addition, research outside of academia shows that consumers' choices are more likely to converge when preference of others is provided ([20, 13]). In contrast, other research has shown that the emergence of online markets based on recommender systems promotes sales of commodities that would not have been found without the new technology. Brynjolfsson et al. [5] argues that consumers' usage of internet search and discovery tools are associated with the increase of collective sales in niche products by comparing online and offline clothing market. Similarly, Zentner et al. [26] also find that the effect of information technology boosts the sales of niche products in movie consumption.

Both features (convergence and divergence) of new technology may be reflected in new search engines – finding papers that would not have been found otherwise and citing papers that everyone has already cited. In this study, we investigate how this new technology could be affecting the convergence and divergence of citations in the fields of sociology and social work.

3 Data and methodology

3.1 Data Source and Coverage

For this study, we used Clarivate Analytics' Web of Science (WoS) data for calculating citation counts to/from papers and to/from journals. This includes the Science Citation Index Expanded, the Social Sciences Citation Index and the Arts and Humanities Citation Index. The full data set includes more than

100 million publications and over 1 billion links between papers from 1900 - 2015. For this study, we isolated the paper counts and citation counts to those papers from the two categories, "Sociology" and "Social Work". The citation counts by year can be found in the appendix.

Since our main research question revolves around the effect of technology on patterns of citation, it is necessary to control other factors such as citation norms [12]. In this article we focus on two relatively similar disciplines in social sciences: Sociology and Social Work, and limit our analysis to within-discipline citation.[3] In the Web of Science Core Collection database, all journals are assigned to one or more subject classifications. While most journals are classified into only one subject category, when a journal is classified in two subject categories we use the first one. In some disciplines subject categorization changes over time; however, classification of journals was relatively stable in Social Work(Last column in Table 5). Sociology experienced two big leaps of journal counts in 2004 and 2008-2009 (Last column in Table 1). The number of papers in both disciplines begins to soar after 2005 and more than doubled by 2013.[4] We believe this rapid increase in publication counts has two sources. First, some of the increase represent a real increase in the number of papers published over time. At the same time, some of the increase is also likely due to an increase in WoS coverage of social science disciplines.[5] It is not clear which factor is more influential in explaining the rapid increase of publication counts after 2005, thus we proceeded with all papers available in the data.

3.2 Data Structures

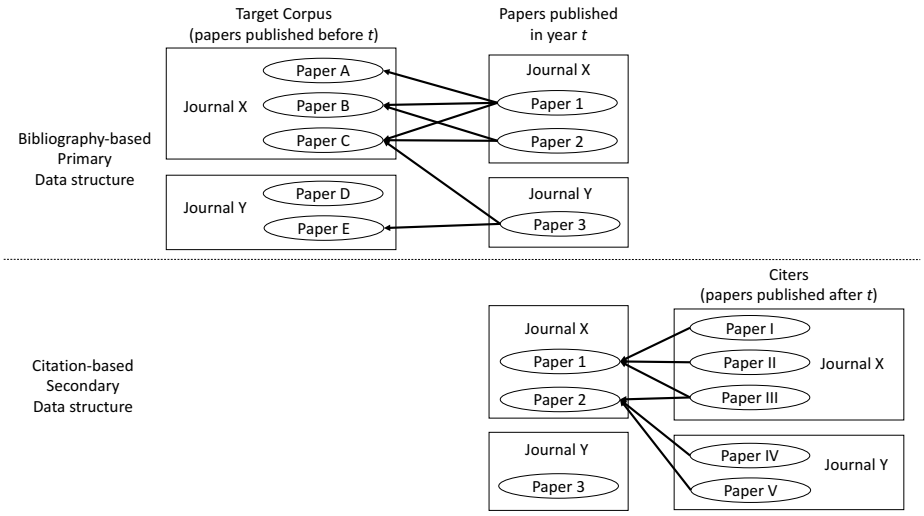
There are several ways to organize bibliographic data for analysis. Two of the most frequently used data structures are a bibliography-based data structure and a citation-based data structure. A bibliography-based data structure focuses on outgoing ties, in the sense that it identifies all papers contained in the bibliographies of a set or sample of papers; analyses typically focus on the distribution and characteristics of the papers cited in these bibliographies. A citation-based data structure is organized around in-coming ties, and selects a set of papers and the papers that cite them. In both cases, the challenge is to identify the appropriate pool of potential targets (in the bibliographic case) or senders (in the citation case). Lynn [17] solved this problem by identifying articles published in well-known journals in selected disciplines between 1985 and 1986, and counting the citations these paper received in the 20 years following their publication.

[3] We will add more disciplines applying more technology in their research such as computer science or physics to compare the amount of change that they experience.

[4] The small decrease in the number of papers published in 2014 is due to the fact that our data ends in mid 2014.

[5] We are investigating the relative import of each of these sources of increase, and will address this in more detail in subsequent drafts.

Figure 1: Data structures



In order to investigate whether researchers' citations have become more expansive or more concentrated since the rise of academic recommender systems, our data structure must meet two criteria: it must include an appropriate pool of papers that *could have been cited* by scholars working at a particular moment in time, and it must allow us to compare citation behavior over time. Thus while Lynn's approach specifies a target pool of papers that could be potentially cited, it is not suitable for our research question because it does not allow us to easily examine changes in researchers' behavior over time.

We proceed by creating two separate data structures, both of which simultaneously identify a pool of papers that could have been cited and allow us to determine if there are temporal changes in researchers' citation behavior. For the majority of our analyses, we use a data structure defined by the complete list of articles published in year t (where t ranges from 1998 to 2014) and their out-going citations. Descriptive statistics of these anchoring papers and their citations are included in Appendix: Table 1 for Sociology and Table 5 for Social Work. This data structure also includes, for each year t, a target corpus containing all articles published in the same discipline in the prior ten years (and indexed in WoS).[6] For example, for articles published in Sociology in 1998, the target corpus contains articles published in sociology journals between 1988 and 1997; for Sociology articles published in 2014, the target corpus becomes articles published in Sociology journals between 2004 and 2013. These target corpora thus specify defined pools of arguably relevant articles that could have been cited by the articles published in a given year. Descriptive statistics for the target

[6]We use a 10 year window because our preliminary analysis showed that across many disciples, most papers cited are between 5 and 10 years old, and that the possibility of citing older papers is relatively low.

corpora are attached in Appendix: Table 3 for Sociology and Table 7 for Social Work. Linking outgoing citations from our lists of anchor articles to the associated target corpora gives us the network of within-discipline out-going citations for a given publication year.[7] Temporal changes in the pattern of citations to the target corpora may reflect changes in scholars' use of technology.

In addition to this primary data structure, we also create a second data structure that allows us to approach the problem from the opposite perspective: what is the pattern of incoming citations received by a defined set of articles published in a given year? In order to maximize our temporal analyses, in this data structure we limit the window of possible citation to two years, and examine the incoming citations to papers published between 1998 and 2011. As for our prior data structure, we restrict our target papers (in this case, potentially citing papers) to papers indexed in WoS, and published in English, but in any disciplines. Analyses using this data structure are viewed primarily as a check on the robustness our primary analyses and done only in sociology.[8]

3.3 Methodology

We begin by investigating the simple question of whether citation patterns have become more- or less- concentrated over time, and whether changes in the these patterns are temporally proximate to technological changes in how scholars access the literature. Using our primary data structure for each year and discipline, we compute a Gini coefficient for the distribution of citations received by papers in the appropriate target corpus. The Gini coefficient ranges from 0 to 1, with values closer to zero reflecting more equal distributions and values closer to one reflecting more unequal (or concentrated) distributions. The Gini coefficient has been used frequently in bibliometrics to evaluate citation distribution [6, 4]. In our context, when a relatively small number of articles receive the bulk of the citations, the distribution of citations within the target corpus will be unequal and the Gini coefficient will be high, while if citations are distributed more evenly across a larger group of target articles, there will be less inequality and the Gini coefficient will be lower. While the Gini coefficient is not a perfect measure of the details of a distribution, it is a powerful and commonly used single number that summarizes the level of relative inequality.

Next, in an effort to better understand factors that impact citation behavior, we estimate a statistical model that predicts the citation counts of papers in each target corpus. Our primary interest here is in whether, in the wake of

[7] The percentage of citations made to the applicable target corpora is stable and low, ranging between 4 and 7 percent in both disciplines. Detailed analysis of citations outside the target corpora are provided in Appendix: Table 2 for Sociology and Table 6 for Social Work. Briefly, about half of all citations go to sources such as news articles, datasets, books, and internet sources that are not indexed by the WoS database. Among those citations to sources that are indexed in WoS, about 40% are to papers that are outside our 10-year window. In addition to these conditions, we also limit our target corpora to article format (which does not include conference proceedings or book reviews) written in English , and in the same discipline.

[8] We plan to extend it to other disciplines as well.

new search technologies, there has been a decline in the impact of the journal a paper is published in and an increase in the impact of the papers' previous record of citations on the predicted number of citations that a paper will receive in a given year.

Our dependent variable, the number of citations received by paper j in year t, contains a lot of zeros (see the last column in Table 3 and Table 7), so we specify a zero-inflated negative binomial model. Because articles are nested in journals, we control this group-level variance with a random effect for journal. Our two main explanatory variables are the journal influence factor[9] (JIF) and the cumulative previous citations received by paper j in years prior to year t. Following convention, JIF is measured by the average citations to articles published in a journal in the two years following publication. This measure, which includes all received citations regardless of disciplines, has previously been computed for all years and all journals in WoS by one of the authors.[10] JIFs are recalculated for each year, though empirically they are relatively stable year over year. We use the year t influence factor of the journal paper j was published in (rather than the publication year) because we are modeling factors that influence the behavior of scholars making decisions in year t about what literature to cite. The cumulative previous citation is the total number of within-discipline citations to paper j through time *t-1*. Following Evans [9] and Lynn [17], we include three control variables, all measured on paper j: article age, page count, and the number of references in the paper's bibliography. Based on our expectation in the introduction, we hypothesize that the effect of JIF has decreased while the effect of cumulative previous citations has increased in recent time periods as the academic recommender system has been popularized.

Our final set of analyses uses our prospective data set to further investigate whether there has been a change in the importance of journals as curators of the scientific literature. In these analyses we classify journals into tiers based on their JIF (top, middle, bottom) and examine temporal changes in the distribution of subsequent citations to papers published in journals in different tiers. Here we are testing whether there has been an increase in the rate of citation for articles published in lower tiered journals, and a decline in the fraction of papers cited in higher tier journals.

[9]We use the ArticleInfluence score for our journal influence measure. This is a network-based journal-level method for ranking journals[24]. It is based on the Eigenfactor algorithm and normalized by the size of the journal.

[10]However, this measure has a limit in that it ignores the distribution of citation across papers within a journal, and thus can be influenced by a single paper that manages to garner a large number of citations. See Milojevic, Radicchi, and Bar-Ilan 2017 [18] for an alternate measure that incorporates within-journal variation in received citations.

4 Results

4.1 Time trend in Gini coefficients[11]

Figure 2 summarizes the Gini coefficients computed from papers that have been cited *at least once* from Sociology and Social Work. Higher Gini values indicate less parity, with a higher concentration of citation wealth to a small proportion of papers. The 95% confidence intervals are generated with 1000 simulations that randomly sample papers from the given citation distribution in year t with replacement. We calculated this for a 3 and 5 year window and found similar results. The Gini coefficients have increased over time, with a particularly steep rise after 2005 in both disciplines. This indicates that citations have concentrated on a few star papers.[12] This finding is consistent with previous observations. For example, multiple studies have shown a Google scholar effect, where more citations go to old and popular articles [23, 21], which implies that citations concentrate on a few popular papers.

Figure 2: Gini coefficients for papers cited at least once between 1998 and 2014. The shaded regions indicate 95% confidence intervals. Blue line represents Sociology and red line represents Social Work.

However, analyzing papers only papers that are cited might overlook the ef-

[11] We are continuing to do more robustness checks on our all findings to investigate whether this tendency is driven by actual behavior change of scholars or a natural outcome of publication and citation count increase.

[12] We argue that this trend is independent from the rapid rise of the number of publications since 2005, because Gini coefficient is a relative inequality measure that accounts for the total number of citations.

fect of technology in increasing access to the broader literature, including those papers from lower tier journals. Figure 3 shows the opposite pattern from Figure 2; when we include all paper including those that have never been cited, the Gini coefficient decrease over time. This suggests a more equal distribution of citations. We call this decrease in Gini coefficients 'divergent'. In these two fields, the distribution of citations has been diverging since early 2000. Interestingly, the divergent trend begins to stagnate after 2010 in both disciplines. If technology had some effect on the divergence since 2000, what has been happening since 2010? Technogical change may be having less of an effect to due broad adoption. We plan to further investigate this stagnation.

Figure 3: Gini coefficients with all papers in corpus between 1998 and 2014. The shaded regions indicate 95% confidence intervals. Blue line represents Sociology and red line represents Social Work.

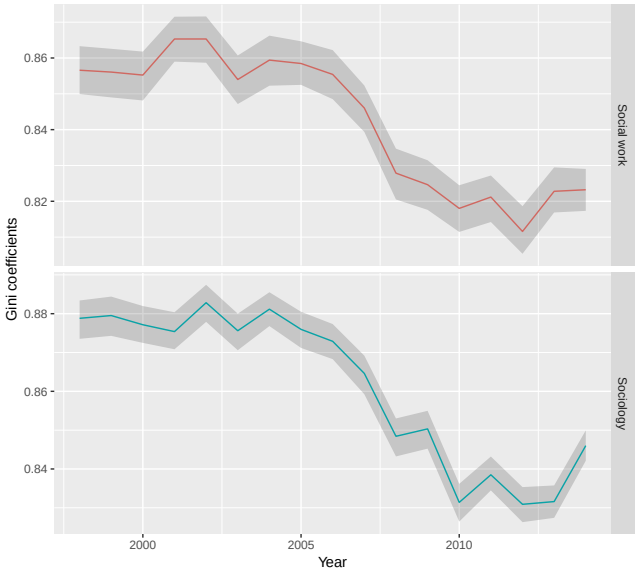
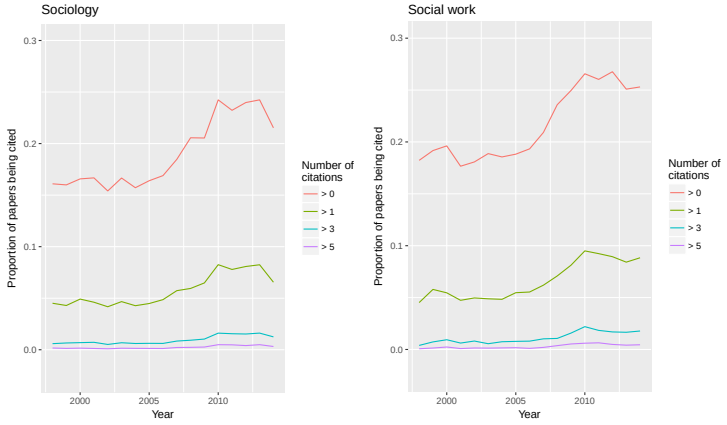


Figure 4 partially explains the contrasting results Figure 2 (increases in the Gini coefficients) and Figure 3 (decreases in the Gini coefficients). Figure 4 shows the percentage of papers that are cited more than 0, 1, 3, and 5 times between 1998 and 2014. The percentage of papers cited at least once and at least twice was stable until about 2005 in both disciplines, but began to rapidly increase after 2005. The groups of papers cited more than three times has not been increasing at the same rate. In other words, about 50% more papers are cited at least once in 2010 than 2005. This increase seems to be driven mostly by the expansion of once or twice cited papers rather than the highest cited papers. We hypothesize that this change may be driven partly by digitization and new search technologies for accessing this portion of the literature.

Figure 4: Proportion of cited papers for Sociology and Social Work. The red line indicates the proportion of papers that have been cited at least once. The purple line is the proportion of papers cited more than five times. The largest increase occurs with the group cited at least once.



4.2 The influence of journal influence factor and previous citations

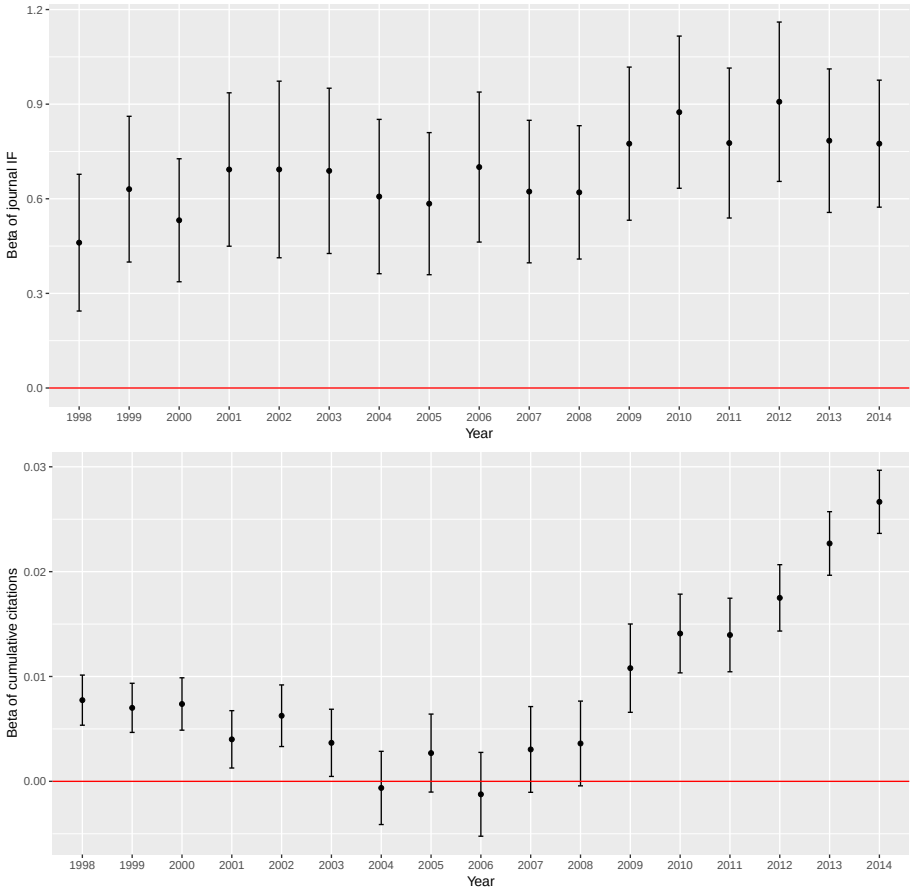
Through analyzing Gini coefficients over time, we find that citation distributions show more divergent rather than convergent behavior. One possible explanation is that technology might induce the citation distribution to be more divergent rather than convergent by facilitating search for papers that were not easily accessible without technology. However, this analysis leaves two related questions. First, why does the potential role of technology in expanding search result in stagnation after 2010? Second, is the reason for Gini coefficients becoming stable after 2010 due to a cancelling out effect on the concentration of citations driven by popularized academic recommender system? In this section, we explore whether the academic recommender system influences researchers to cite papers that have high previous citation counts rather than high status journals by comparing the magnitude of coefficients of the two variables computed from the statistical model, JIF and paper's previous cumulative citations, between 1998 and 2014. All 17 models have Variance inflation factor specified for this statistical model less than five, which suggests that there are no issues of collinearity.[13]

Figure 5 and Figure 6 illustrate the results of the negative binomial zero-inflated random effect model.[14] In both figures, the top panel summarizes JIF coefficients and the lower panel shows cumulative previous citations. Also, 95% confidence intervals are added to the estimated coefficients. In sociology results (Figure 5), while the estimated coefficients of JIF are around 0.3 and 0.9 and

[13] Correlation tables are attached in Appendices (Sociology for 4 and Social Work for 8).
[14] 34 full models will be provided as a supplementary document in future.

generally higher than cumulative previous citations, it is hard to argue that the coefficients of JIF have changed during observed period due to its wide confidence interval. In contrast, the increase in cumulative previous citations stands out since 2009. Between 2003 and 2008, cumulative previous citations are not statistically significant with .05 alpha level, but after 2009, it is statistically significant and the predicted coefficients have constantly increased. The increased coefficients after 2009 indicate that the effect of one citation increase in previous citation counts is associated with more received citations by paper in year t.

Figure 5: Coefficients of journal influence factor and cumulative previous citations (Sociology, 95% C.I)



Social Work results (Figure 6[15]) show a similar pattern with sociology; while coefficients of JIF do not show statistically significant change, coefficients of cu-

[15]The result of 2005 for Social Work is not included because the model fails to converge.

mulative previous citations begin to increase after 2007. Consistently increasing effect of previous citation count in both disciplines after controlling JIF demonstrates that the paper-based measure now has separate meanings from JIF in predicting received citations. The increased trend cannot be attributed to changes in average cumulative citation count in corpora and average received citations by papers in year t. First, averages have constantly fluctuated since 1998. The effect of previous citation count only increases in the most recent period. Also, coefficients are associated with variance of independent and dependent variables, thus the increase in averages do not necessarily lead to the increase in coefficients.

Figure 6: Coefficients of journal influence factor and cumulative previous citations (Social Work, 95% C.I)

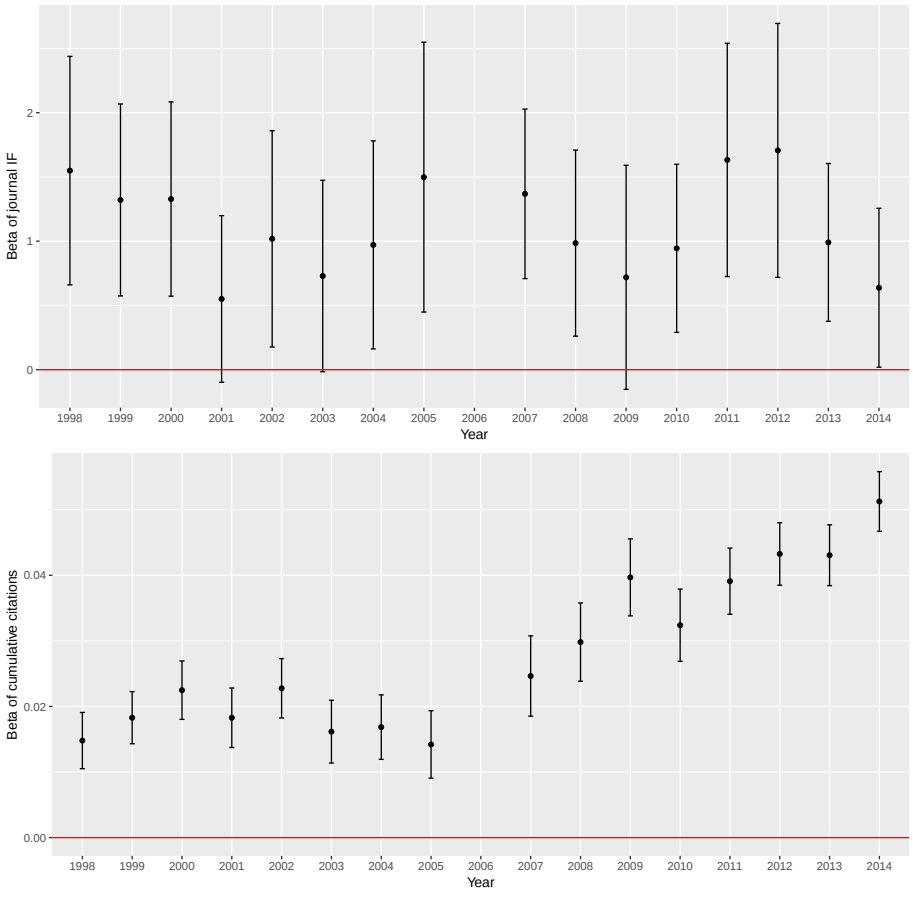
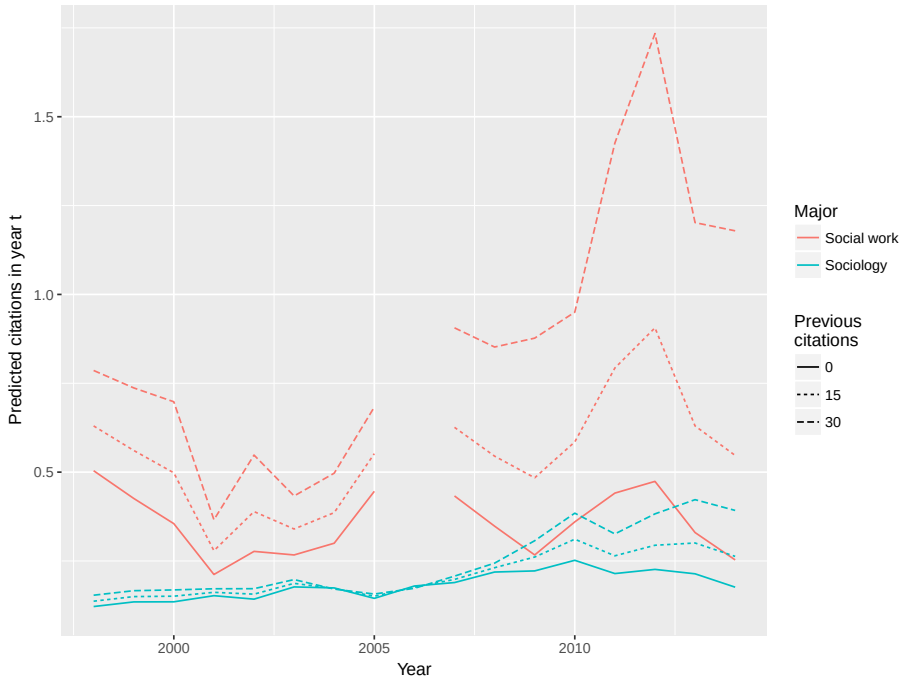


Figure 7[16] provides the predicted number of being cited in year t when the number of previous cumulative citations is 0, 15, or 30 after controlling other variables. Other variables are set to age 5, published in a journal with influence factor 1, and the average number of references and page counts in each discipline. By simulating expected number of being cited based on the same hypothetical scenarios, the figure helps understand the meaning of increase in coefficients of previous citation counts. In the figure, red lines represent Social Work and blue lines represent sociology. In both disciplines, the influence of previous cumulative citations becomes more significant since 2008 than before. Particularly, in sociology, previous cumulative citations do not make any differences in predicted outcome; however, in 2014, an article cited 30 times is expected to receive .25 more citations than an article never cited. The difference is larger in Social Work, in 2014, an article cited 30 times is expected to be cited 1 time more than an article never cited.

Figure 7: Predicted number of being cited in year t



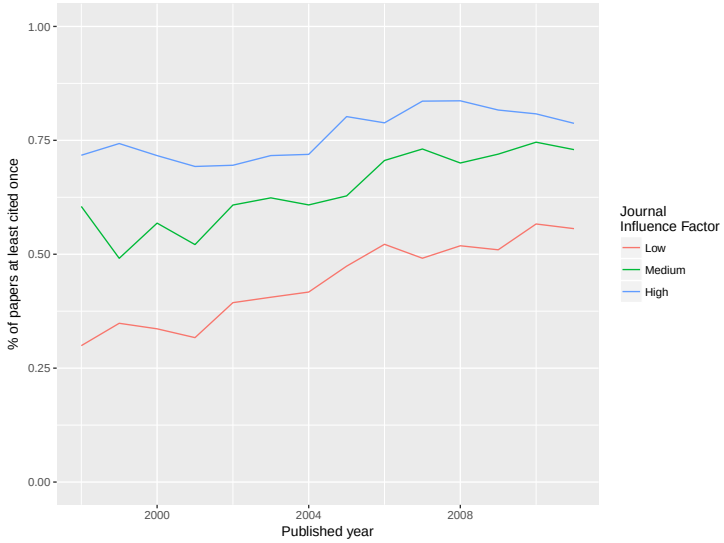
[16]This figure lacks the predicted number of being cited in year t since the model for Social Work in 2005 fails to converge.

4.3 Journal variability

Until now, we illustrate two main findings: first, the distribution of citations has been more equal after 2005 than before; second, the importance of the measure of paper-based popularity, previous cumulative citations, has been increased in explaining the number of received citations in time *t*. What two findings imply is that the development of search technology enables the articles - published in low tier journals and thus not broadly exposed to researchers - to be now searched if a research topic and typed keywords are well matched. However, results above do not provide direct evidence that the exposure of low tier journals has broadened. Therefore, in this section, we seek whether the role of journal status has changed in predicting incoming citations by using the second citation-based data structure.

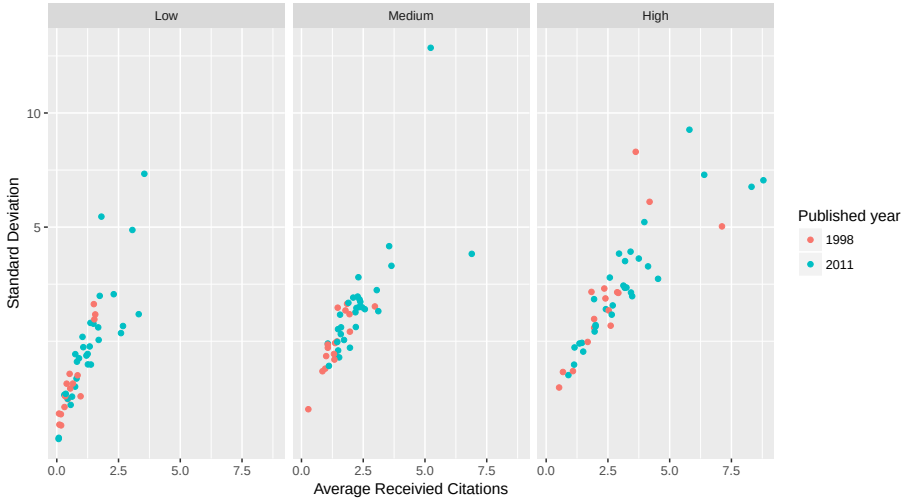
Figure 8 summarizes the percentage of papers cited at least once between 1998 and 2001. We sort journals by JIF and divide it into three equal parts. The blue line in the figure represents high JIF group, green means medium, and red means low. While the percentage of papers at least cited once in high JIF group stays around 75% in all years, both medium and low groups experience a rapid increase during the time period. This trend indicates that not only articles published in high tier journals, but others in medium or low tier journals also become more cited often in recent years, which implies that the advancement of technology might help articles not published in top journals be easily exposed to researchers. While journals are still an important signal of the quality of a paper, it is now possible for the papers to have a chance to be read and cited regardless of their JIF.

Figure 8: Percentage of papers cited at least once in two years after published between 1998 and 2011 (Sociology)



However, Figure 8 has a limit because it shows how many papers have a chance to be cited at least once, but does not explain whether received citations are rather concentrated on a few star papers in journal or equally distributed. Figure 9 suggests one way to answer this question by plotting average received citations and standard deviation of a journal in 1998 and 2011. We take a square root in y-axis to simplify the graph. While it is not possible to assume that publications from two years are in the exact same condition because the average received citations has increased following the increase of publications in recent years, there are more number of journals with very high standard deviation in low or medium tier journals in 2011 than 1998.[17] A few journals in low or medium tier journals now have high variance of citations, which means that journal status might not be a decisive factor in predicting popularity of published papers.

Figure 9: Standard deviation of the number of cited in two years after published (Sociology)



5 Discussion and conclusion

When preparing this manuscript, we used various search engines, including Google Scholar, to find appropriate citations – the very tools we aim to analyze in this paper. Sometimes we had a specific paper in mind; other times we had to find papers to either support or refute a statement in our paper. Very rarely did we move beyond the first page. If other researchers are following similar search behaviors, the effect on what is searched and found in the literature,

[17] While we compare only two years for simplicity, this pattern is consistent over time; there are no journals having standard deviation greater than 4 before 2003 in low and medium tier.

we claim, is worthy of investigation.

How we search papers is dramatically different than how we searched papers just a couple decades ago. If we know the title or DOI, it is generally not difficult to download the PDF within minutes. In this study, this is not the kind of search we are investigating. Instead, we are trying to access the effect of new digital technologies on literature we don't know exists – when searching new disciplines, concepts, or appropriate citations. Are these technologies helping us, collectively, find more or less of the relevant literature? Our objetive in this paper is to investigate the effect of digital technology, academic search engines and recommender systems on what is being searched, people's academic careers, and novel discovery.

To investigate these questions, we start by looking at citation patterns in the fields of Sociology and Social work. We find that more of the literature is being cited (i.e., less zero cited papers) but higher concentration of citations towards the star papers. This is consistent with what others have found [7, 1]. We cannot answer whether these patterns are being strictly driven by academic search engines, but they are suggestive given the introduction of these technologies. More work will need to be done for teasing out these relationships. Other factors could be driving these patterns. For example, conferences, twitter, facebook and related technologies may be highlighting papers forgotten in previous decades.

In our results, we find that the impact of previous cumulative citations representing the influence of academic recommender system has been up from around 2008 to 2014. The similar pattern has been observed from the second set of analysis; recently, more papers published in low- or medium-tier journals are cited at least once in two years and the variance of the number of received citations among papers published in the same journal is generally higher in 2014 than 1998. This finding is consistent with how Google Scholar is designed, which suggests that more scholars are using a sort of new search engines in making bibliographies of their research. While all papers are prone to be exposed depending on keywords they have mostly in title, which was impossible when people walked into the library and picked up renowned journals, but at the same time, when two papers include the same keywords, previously highly cited papers tend to be up first in search page.

Figure 3 shows an interesting pattern as well. Since 2000, the citation distribution has diverged – a higher proportion of papers have received at least one citation. If new digital technologies have been driving this divergence what is happening since 2010? Since then, the decrease has stabilized. We plan to look into this in more detail. It could be that these technologies have been broadly adopted. Therefore, the changes in citation distribution are less dramatic.

A limitation of this study is that we might ignore classic papers in the analysis such as works written by Karl Marx or Thomas Kuhn that have explosive influential power. The preferential attachment mechanism might be a better explanation if we only look at papers that have received several thousand of citations. However, we decide to exclude them for two reasons. First, as we explained in the data structures, it is necessary to identify a pool of papers that would have been cited; if we decided to include classic papers, a pool of papers

would have covered publications from early 20th century. Second, the intention of researchers citing recent papers and classic foundational ones might be different; researchers might cite more foundational papers when they need to bring authority in their research to persuade readers to emphasize the significance of the study ([12]). Thus, we believed that it is more important to separate two different kinds of citations and isolate the effect of technology than include star papers to answer our main research questions.

We do not make any judgements whether the transforming behavior of researchers is beneficial in sustaining the healthy academic environment or vice versa. The literature roughly doubles every 20 years, and as this expansion continues, it will be more and more difficult for scholars to keep up with even their own fields. Recommendation algorithms will be needed for assisting scholars in the literature searches. However, we argue that it is important to understand and monitor how the recommender system changes the way of doing research, particularly, as these recommendation algorithms become more come for every day research. Based on our findings so far, the new technology does not passively assist researchers' job in searching literature, but may actively interfere in researchers' evaluation of which papers are more important to be cited for their scientific work.

While we mainly argue that the development of technological development in search engine drives transformation of researchers' behavior and it is suggested by changing citation distributions over time, this change cannot be solely attributed to the effect of technology. There are other possible scenarios that might lead to researchers' behavior change. First, self-citations made by authors as well as journals might have been recently increased, which helps increase the percentage of papers at least cited once in low- or medium-tier journals. As more journals and researchers are evaluated their performance by the number of citations that they receive, editors of journal and researchers might try more to cite papers from submitting journals or themselves in a working article. Second, topics inside of disciplines might have been specialized, which leads to researchers to cite inside of a specialized discipline regardless of journal prestige. It might also broaden the percentage of papers ever cited from non-top, but specialized journals. We could not examine these possibilities in our current analyses, but we plan to investigate this as a next step.

6 Acknowledgments

We want to thank Clarivate Analytics for providing the Web of Science data.

References

[1] Réka Albert and Albert-László Barabási. Topology of evolving networks: local events and universality. *Physical review letters*, 85(24):5234, 2000.

[2] Joeran Beel, Stefan Langer, Marcel Genzmehr, Bela Gipp, Corinna Breitinger, and Andreas Nürnberger. Research paper recommender system evaluation: a quantitative literature survey. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, pages 15–22. ACM, 2013.

[3] Jöran Beel and Bela Gipp. Google scholar's ranking algorithm: an introductory overview. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, volume 1, pages 230–241. Rio de Janeiro (Brazil), 2009.

[4] Lutz Bornmann, Rüdiger Mutz, Christoph Neuhaus, and Hans-Dieter Daniel. Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in science and environmental politics*, 8(1):93–102, 2008.

[5] Erik Brynjolfsson, Yu Hu, and Duncan Simester. Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science*, 57(8):1373–1386, 2011.

[6] Quentin L Burrell. The bradford distribution and the gini index. *Scientometrics*, 21(2):181–194, 1991.

[7] D. J. de Solla Price. Networks of scientific papers. *Science*, 149:510–515, 1965.

[8] Anita Elberse. Should you invest in the long tail? *Harvard business review*, 86(7/8):88, 2008.

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

[10] Daniel Fleder and Kartik Hosanagar. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science*, 55(5):697–712, 2009.

[11] Jim Giles. Science in the web age: Start your engines. *Nature*, 438(7068):554–555, 2005.

[12] Lowell L Hargens. Using the literature: Reference networks, reference contexts, and the social structure of scholarship. *American sociological review*, pages 846–865, 2000.

[13] Tad Hogg and Kristina Lerman. Disentangling the effects of social signals. *arXiv preprint arXiv:1410.6744*, 2014.

[14] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431, 2015.

[15] Vincent Lariviere, Veronique Kiermer, Catriona J MacCallum, Marcia McNutt, Mark Patterson, Bernd Pulverer, Sowmya Swaminathan, Stuart Taylor, and Stephen Curry. A simple proposal for the publication of journal citation distributions. *Biorxiv*, page 062109, 2016.

[16] Kristina Lerman and Tad Hogg. Leveraging position bias to improve peer recommendation. *PloS one*, 9(6):e98914, 2014.

[17] Freda B Lynn. Diffusing through disciplines: Insiders, outsiders, and socially influenced citation behavior. *Social Forces*, 93(1):355–382, 2014.

[18] Staša Milojević, Filippo Radicchi, and Judit Bar-Ilan. Citation success index an intuitive pair-wise journal comparison metric. *Journal of Informetrics*, 11(1):223 – 231, 2017.

[19] Raj K Pan, Alexander M Petersen, Fabio Pammolli, and Santo Fortunato. The memory of science: Inflation, myopia, and the knowledge network. *arXiv preprint arXiv:1607.05606*, 2016.

[20] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

[21] Alexander Serenko and John Dumay. Citation classics published in knowledge management journals. part ii: studying research trends and discovering the google scholar effect. *Journal of Knowledge Management*, 19(6):1335–1355, 2015.

[22] Daniel Torres-Salinas, Álvaro Cabezas-Clavijo, and Evaristo Jiménez-Contreras. Altmetrics: New indicators for scientific communication in web 2.0. *arXiv preprint arXiv:1306.6595*, 2013.

[23] Alex Verstak, Anurag Acharya, Helder Suzuki, Sean Henderson, Mikhail Iakhiaev, Cliff Chiung Yu Lin, and Namit Shetty. On the shoulders of giants: The growing impact of older articles. *arXiv preprint arXiv:1411.0275*, 2014.

[24] J.D. West, T.C. Bergstrom, and C.T. Bergstrom. The eigenfactor metrics: A network approach to assessing scholarly journals. *College and Research Libraries*, 71(3):236–244, 2010.

[25] Jevin D West, Michael C Jensen, Ralph J Dandrea, Gregory J Gordon, and Carl T Bergstrom. Author-level eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *Journal of the American Society for Information Science and Technology*, 64(4):787–801, 2013.

[26] Alejandro Zentner, Michael Smith, and Cuneyd Kaya. How video rental patterns change as consumers move online. *Management Science*, 59(11):2622–2634, 2013.

Appendices

A Descriptive statistics (Sociology)

Year	Total papers published	Total citations made	Average citations	Cites to corpus	Percent of cites to corpus	Total journals
1998	1577	62280	42	3743	6.0%	90
1999	1631	63712	42	3674	5.8%	91
2000	1744	70369	42	3920	5.6%	91
2001	1680	67862	42	3904	5.8%	91
2002	1620	69074	45	3533	5.1%	92
2003	1699	72842	44	3936	5.4%	92
2004	1674	72802	46	3679	5.0%	99
2005	1763	78314	45	3853	4.9%	101
2006	1963	90863	47	4073	4.5%	102
2007	2185	99823	47	4752	4.8%	104
2008	2658	121584	47	5367	4.4%	111
2009	2812	132025	48	5854	4.4%	127
2010	3296	165696	51	7885	4.8%	129
2011	3569	185759	53	8123	4.4%	129
2012	3781	193628	52	8988	4.6%	128
2013	4065	208501	52	10118	4.8%	127
2014	3635	189883	53	9296	4.9%	123

Table 1: Bibliography (Sociology)

Year	Total citations	WOS (%)	WOS + 10-year (%)	WOS + 10-year + English + Article (%) (Will be added)	WOS + 10-year + English + Article + Sociology (%)
1998	65079	62.7	39.6		6.0
1999	68457	61.7	39.4		5.8
2000	72938	61.7	38.2		5.6
2001	70145	62.1	37.2		5.8
2002	70602	60.8	36.2		5.1
2003	74622	60.6	36.6		5.4
2004	74695	60.8	35.9		5.0
2005	82045	60.4	36.2		4.9
2006	92361	59.7	35.3		4.5
2007	102595	61.2	37.1		4.8
2008	124669	59.2	35.5		4.4
2009	134046	58.4	35.3		4.4
2010	168143	57.4	34.4		4.8
2011	189040	53.1	30.4		4.4
2012	195472	38.2	22.2		4.6
2013	210669	39.6	22.7		4.8
2014	193204	41.2	23.5		4.9

Table 2: Origin of citations (Sociology)

Year	Year window	Total papers in corpus	Total received citations	Papers cited at least once	Percentage of papers cited at least once
1998	1988-1997	16072	3743	2587	16%
1999	1989-1998	15976	3674	2555	16%
2000	1990-1999	15910	3920	2637	17%
2001	1991-2000	16074	3904	2680	17%
2002	1992-2001	16105	3533	2481	15%
2003	1993-2002	16171	3936	2694	17%
2004	1994-2003	16312	3679	2564	16%
2005	1995-2004	16389	3853	2687	16%
2006	1996-2005	16581	4073	2801	17%
2007	1997-2006	16977	4752	3135	18%
2008	1998-2007	17536	5367	3606	21%
2009	1999-2008	18617	5854	3824	21%
2010	2000-2009	19798	7885	4799	24%
2011	2001-2010	21350	8123	4959	23%
2012	2002-2011	23239	8988	5573	24%
2013	2003-2012	25400	10118	6157	24%
2014	2004-2013	27766	9296	5977	22%

Table 3: Target corpus (Sociology)

	Age	Page count	Reference count	Previous cumulative citations	Journal influence factor
Age	1.000	-0.062	0.019	0.130	0.054
Page count		1.000	0.487	0.345	0.157
Reference count			1.000	0.626	0.228
Previous cumulative citations				1.000	0.054
Journal influence factor					1.000

Table 4: Correlation of variables in 1998 (Sociology)

B Descriptive statistics (Social Work)

Year	Total papers published	Total citations made	Average citations	Cites to corpus	Percent of cites to corpus	Total journals
1998	932	31364	34	2083	6.6%	36
1999	955	33838	36	2434	7.2%	36
2000	943	34548	37	2569	7.4%	36
2001	893	31908	36	2269	7.1%	36
2002	916	33462	37	2402	7.2%	36
2003	893	33223	37	2403	7.2%	36
2004	928	36731	40	2441	6.7%	34
2005	981	38859	40	2560	6.6%	35
2006	1106	42300	39	2643	6.2%	36
2007	1222	48427	40	2967	6.1%	36
2008	1344	55064	41	3484	6.3%	39
2009	1553	62728	41	4084	6.5%	42
2010	1694	74537	44	4873	6.5%	44
2011	1721	77208	45	5048	6.5%	42
2012	1771	80376	45	5344	6.7%	43
2013	1790	81789	46	5338	6.5%	41
2014	1691	78779	47	5876	7.5%	41

Table 5: Bibliography (Social Work)

Year	Total citations	WOS (%)	WOS + 10-year (%)	WOS + 10-year + English + Article (%) (Will be added)	WOS + 10-year + English + Article (%) + Social Work
1998	31453	72.5	50.8		6.6
1999	34125	72.9	50.2		7.2
2000	35150	71.0	48.1		7.4
2001	31941	70.8	47.9		7.1
2002	33482	70.0	47.4		7.2
2003	33653	69.9	46.3		7.2
2004	36892	69.1	45.6		6.7
2005	39353	69.6	45.9		6.6
2006	42752	68.7	45.9		6.2
2007	48442	68.8	46.1		6.1
2008	55377	67.0	43.7		6.3
2009	63109	66.9	44.6		6.5
2010	75593	65.6	43.4		6.5
2011	77866	61.0	39.4		6.5
2012	80470	45.9	29.0		6.7
2013	81927	47.9	30.0		6.5
2014	79740	49.0	30.5		7.5

Table 6: Origin of citations (Social Work)

Year	Year window	Total papers in corpus	Total received citations	Papers cited at least once	Percentage of papers cited at least once
1998	1988-1997	8428	2083	1536	18%
1999	1989-1998	8639	2434	1656	19%
2000	1990-1999	8848	2569	1736	20%
2001	1991-2000	9054	2269	1599	18%
2002	1992-2001	9102	2402	1645	18%
2003	1993-2002	9176	2403	1732	19%
2004	1994-2003	9248	2441	1716	19%
2005	1995-2004	9340	2560	1758	19%
2006	1996-2005	9336	2643	1805	19%
2007	1997-2006	9472	2967	1981	21%
2008	1998-2007	9769	3484	2304	24%
2009	1999-2008	10181	4084	2541	25%
2010	2000-2009	10779	4873	2864	27%
2011	2001-2010	11530	5048	3001	26%
2012	2002-2011	12358	5344	3308	27%
2013	2003-2012	13213	5338	3315	25%
2014	2004-2013	14110	5876	3570	25%

Table 7: Target corpus (Social Work)

	Age	Page count	Reference count	Previous cumulative citations	Journal influence factor
Age	1.000	-0.110	-0.012	0.162	0.127
Page count		1.000	0.294	0.250	-0.111
Reference count			1.000	0.557	0.151
Previous cumulative citations				1.000	-0.032
Journal influence factor					1.000

Table 8: Correlation of variables in 1998 (Social Work)

C GLM (Binomial) results

Figure 10: Coefficients of journal influence of factor and cumulative citations (Sociology, Binomial model)

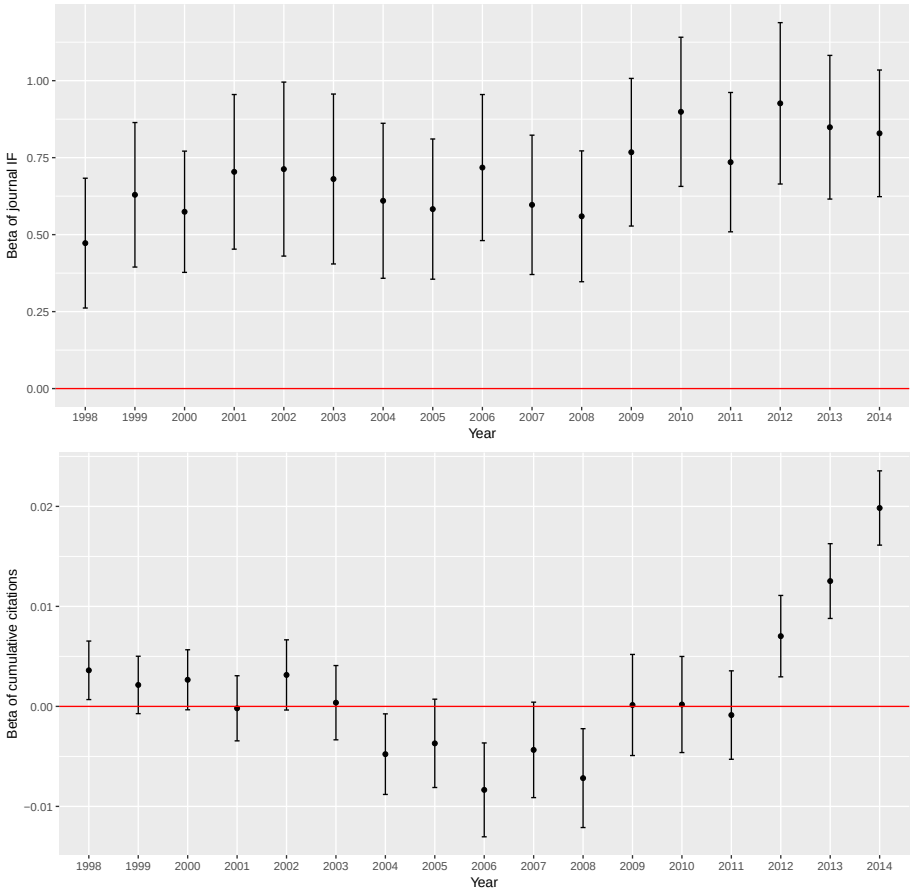


Figure 11: Coefficients of journal influence of factor and cumulative citations (Social Work, Binomial model)

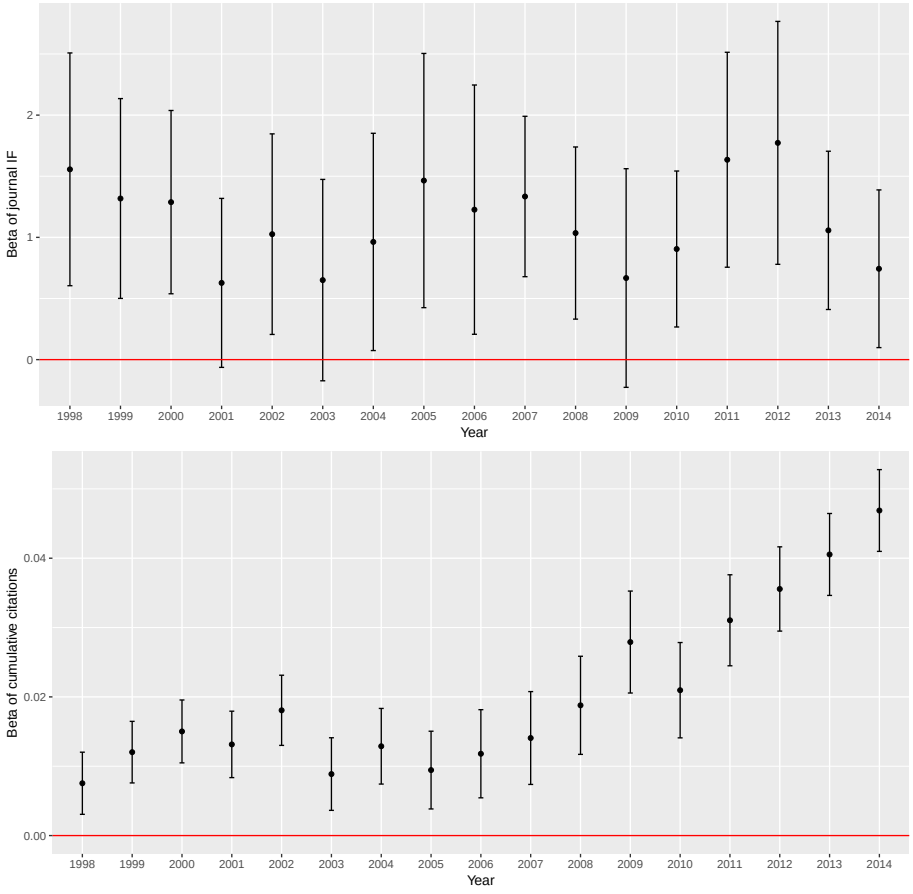


Figure 12: Predicted probability of being cited in year t (Binomial model)