The Influence of Changing Marginals on Measures of Inequality in Scholarly Citations: Evidence of Bias and a Resampling Correction

Lanu Kim★, Christopher Adolph†, Jevin West†, and Katherine Stovel†

★Stanford University, lanu@stanford.edu, *corresponding author*

†University of Washington, Seattle, {cadolph, jevinw, stovel}@uw.edu

forthcoming in *Sociological Science* · 6 July 2020

Abstract. Scholars have debated whether changes in digital environments have led to greater concentration or dispersal of scientific citations, but this debate has paid little attention to how other changes in the publication environment may impact the commonly used measures of inequality. We demonstrate using Monte Carlo experiments that a variety of inequality measures – including the Gini coefficient, the Herfindahl-Hirschman index, and the percentage of papers ever cited – are substantially biased downwards by increases in the total number of papers and citations. We propose and validate a resampling-based correction for this "marginals bias," and apply this correction to empirical data on scholarly citation distributions using Web of Science data covering four broad scientific fields (Health; Humanities; Mathematics and Computer Sciences; and Social Sciences) during 1996–2014. We find that in each field the bulk of the apparent decline in citation inequality in recent years is an artifact of marginals bias, as are most apparent inter-field differences in citation inequality. Researchers using inequality measures to compare citation distributions and other distributions with many cases at or near the zero-bound should interpret these metrics carefully and account for the influence of changing marginals.

Keywords: Citation distribution; Citation pattern; Uncitedness; Concentration; Gini coefficient; Herfindahl-Hirschman Index; Inequality

Acknowledgements: We thank Clarivate Analytics for providing the Web of Science data, and Elena Erosheva, Bas Hofstra, and Joe Cho for helpful conversations.

Declarations of interest: None.

Funding: This research was supported by National Science Foundation grant #1735194, Katherine Stovel PI, Jevin West co-PI.

1 Introduction

While the structure of citations to scholarly papers has been studied since de Solla Price's 1965 seminal work, this line of research has recently been reinvigorated as a result of digitization of publications and new forms of communication and search. These technological developments have lead some optimists to claim that increased access to previous research will enhance exposure to new ideas and stimulate scientific discovery. However, others have worried that algorithmically-driven tools will concentrate scientists' attention on a small number of "star" papers, leading to more derivative and less ground-breaking research.[1] Which tendency dominates will have important implications for the future of scientific advancement (Hamilton, 1990; Larivière, Gingras & Archambault, 2009; Evans, 2008; Barabási, Song & Wang, 2012; Evans & Reimer, 2009; Eysenbach, 2006), largely because of the well-recognized advantages of epistemic diversity on innovation (Zollman, 2010; Weatherall & O'Connor, 2020; Larsen, Hovorka, Dennis & West, 2019; Hofstra, Kulkarni, Galvez, He, Jurafsky & McFarland, 2020). The empirical evidence put forth thus far in studies of the distribution of scientific citations is contradictory. Focusing on the impact of the rise of online journal access, one study found evidence of increasingly concentrated citations (Evans, 2008), while other analyses of aggregate trends over time revealed more diversified citations (Larivière et al., 2009).

We contribute to this discussion by focusing directly on an unrecognized limitation in various inequality measures, including the Gini coefficient and the percentage of ever cited papers, that are commonly used to gauge the level of concentration. Our specific concern is that when the unit to be distributed is indivisible (as are citations) and on roughly the order of magnitude as the number of targets (as are citations and publications), inequality measures are highly sensitive to changes in the input marginals. We investigate this problem in the context of scientific citations, and demonstrate that

1 Studies investigating whether algorithmically-driven online portals concentrate or broaden exposure are not limited to scientific citation behavior, but also include consumer decisions in online clothing markets (Brynjolfsson, Hu & Simester, 2011), video rentals (Zentner, Smith & Kaya, 2013), and music consumption (Salganik, Dodds & Watts, 2006).

marked and uneven growth in the number of publications and citations affects measures of inequality and confounds year-over-year and between-field comparisons.

As Figure 1 shows, in each of four broad disciplines the number of papers published and citations to these papers has increased since 1996, in some cases dramatically.[2] Further, the growth in the two quantities is not proportional, with the number of citations generally increasing more rapidly than the number of publications. This dramatic growth in publications and citations has caught the attention of others who study scientific knowledge production, most notably Wallace et al. 2009 who report that most of the decline in uncitedness between 1900 and 2006 is a result of the increase in subsequent publications (and total citations made by those publications). General discussion of the expansion in publications appears in studies of inflation in journal impact factors and article-based citation measures (Althouse, West, Bergstrom & Bergstrom, 2009; Petersen, Pan, Pammolli & Fortunato, 2019), the aging of the scientific literature (Larivière, Archambault & Gingras, 2008; Parolo, Pan, Ghosh, Huberman, Kaski & Fortunato, 2015), and the growing myopia of science (Pan, Petersen, Pammolli & Fortunato, 2018).

However, there has been no investigation of how these changes in the volume of publication and citation might bias interpretation of the specific measures of inequality typically used to capture how citations are distributed across the scientific literature. Because fully capturing the shape of a distribution with a single number is impossible, many different approaches to measuring inequality have been proposed. One simple approach is to calculate the share of one value or entry in the total distribution, such as the number of papers never cited (Wallace et al., 2009; Fleder & Hosanagar, 2009; Zentner et al., 2013); another approach is to summarize the shape of the distribution with respect to its total deviation from a uniform distribution. The Gini coefficient (Brynjolfsson et al., 2011; Salganik et al., 2006; Varga, 2019) and the Herfindahl-Hirschman index (Evans, 2008) are well-known examples of this latter approach. Each measure of inequality has limitations, most conspicuously that differently shaped distributions may generate the same value (Atkinson, 1975), and the possibility of bias in small samples (Deltas, 2003). Other, less appreciated problems plague their use in studies of scholarly citations: citations to papers are not divisible; the total number of citations is sometimes less than the number of citable papers; and in most fields, large fractions of papers are never cited, mixing large numbers of zeros into citation distributions (Bryn-

2 In the Humanities, the number of citations received is less than the number of articles published, due to the practice of citing books and historical records not indexed in Web of Science.

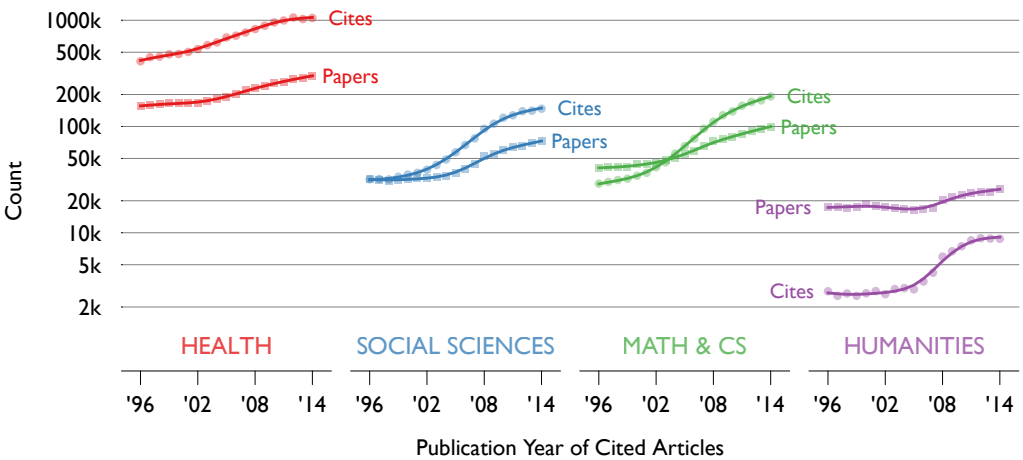

Figure 1. *Number of journal articles published 1996–2014 and citations to those articles within two years of publication.* Compiled from the Web of Science (Clarivate Analytics). See the Supplementary Materials, section S1, for the composition of the four broad categories shown above. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

jolfsson et al., 2011; Wallace et al., 2009; Larivière et al., 2009). Moreover, changes in the marginal number of papers and citations cause the severity of these problems in citation distributions to vary, which renders comparisons across time and across disciplines difficult. Ignoring these issues, scholars studying population-level citation behavior nevertheless use such inequality measures to draw substantive conclusions about changes over time (e.g., Huang, Chang & Chen, 2012; Ranasinghe, Shojaee, Bikdeli, Gupta, Chen, Ross, Masoudi, Spertus, Nallamothu & Krumholz, 2015; Yoon, Yoon, Lee, Baek, Lim, Seo & Yun, 2017).

And yet, if the aim is to understand whether individual scholars' citing behaviors are changing in ways that aggregate to a different macro-level citation structure, we must be confident that changes in measures of the citation distribution reflect changes in individual decisions rather than other contextual shifts. Because the number of published papers and citations have been steadily increasing (Bornmann & Mutz, 2015; Pan et al., 2018; Petersen et al., 2019), the overall volume of papers published and citations made can be treated as largely exogenous with respect to an individual scholar's choice of specific papers to cite. In the case of the structure of scientific citation, dramatic changes in the number of papers published and citations made will lead to substantial year-over-year changes in the theoretically possible levels of concentration or disper-

sion in citations. A simple example illustrates. If there were 1000 papers published in a given year and only 500 citations made to those papers, the theoretical maximum in the percentage of papers cited at least once is 50%, whereas if there were a total of 1000 or more citations made to those same 1000 papers, the theoretical maximum of the percentage of papers cited rises to 100%. Similar, but more subtle versions of this problem arise for other measures of inequality. Taken together, these problems suggest that comparisons based on standard measures of inequality may be inadequate or even misleading when the marginals of the distributions of papers and citations are changing substantially over time.[3]

Using data from the Web of Science, we first demonstrate that inter-year comparisons of common measures of citation inequality are likely to be biased using a series of Monte Carlo experiments on hypothetical populations of papers. These experiments are constructed to hold patterns of inequality fixed across fields and periods, while allowing the total number of papers and citations to follow their empirically observed trends over years and fields. These results reveal that marginal change in publications and citations itself is sufficient to produce dramatic temporal change in inequality measures. Next, we develop a bias-correction for inequality in the presence of changing marginals and show that this correction appears to completely remove the substantial bias created by trends towards higher total publications and citations. Then, we apply this correction to inequality measures of the observed population of citations. Our adjustment reveals that irrespective of field, the large majority of the apparent decline in citation inequality in recent years is an artifact of bias induced by changing marginals. Rather than declining, citation inequality in the Web of Science database appears to be largely stable over recent decades. Finally, we apply the same correction method to reduce marginals bias when making comparisons between broad fields. After adjustment, most inter-field differences in citation inequality are also revealed to be an artifact resulting from differences in the size of fields.

2 Citation data and inequality measures

We analyze publication and citation data for four broad disciplinary fields that were the focus of Larivière et al.(Larivière et al., 2009) – Health; Humanities; Mathematics and Computer Sciences (Math & CS); and Social Sciences – using Web of Science

3 We use the term "marginals" to refer to the total number of published papers in a year and the total number of citations made to all papers published in that year.

(WoS) data provided by Clarivate Analytics.[4] We categorize the four broad disciplinary fields following the National Science Foundation's taxonomy of disciplines created by the Integrated Postsecondary Education Data System survey. (See the Supplementary Materials, section S1 for further details of categorization.) Within each broad set of fields, we include research papers published in English language journals between 1996 and 2014 and exclude editorial comments, books, and other non research articles. Because of uneven coverage during much of the twentieth century, we limit our analyses to papers published between 1996 and 2014.[5] We drop one unusually well-cited 2004 paper in Math & CS[6] as an effort to understand the general temporal pattern in inequality measures. (See Supplementary Materials (section S2.4) for results that include this outlier.)

Generally following Larivière et al's 2009 approach, we construct a data structure that includes papers published between 1996 and 2014 and citations toward those papers using a series of two year moving windows from 1996 and 2016.[7] For example, for all papers published in the Social Sciences in 2014, we identify citations to these papers from other papers published in the Social Sciences until 2016. Table 1 reports the total number of papers and citations in each broad discipline.

Using these data, we focus on four yearly, field-specific measures of citation inequality: the Gini coefficient; the proportion of papers published in a given year that received at least one citation; the proportion of papers needed to account for 20% and 80%

4 WoS includes journals indexed in the Science Citation Index Expanded, the Social Sciences Citation Index, the Arts and Humanities Citation Index, and the Emerging Source Citation Index.

5 To account for the impact of changes in the coverage of journals in the database during our period of study, we performed a robustness check that includes only journals that appear in the database for all years of our study period. Results, presented in the Supplementary Materials (section S2.1) are consistent with findings presented in the main text.

6 This paper is Sudhir Kumar, Koichiro Tamura, and Masatoshi Nei, MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment, *Briefings in Bioinformatics* 5(2), June 2004, 150–163, https://doi.org/10.1093/bib/5.2.150.

7 We performed additional robustness checks using four-year and six-year citation windows, results of which are provided in the Supplementary Materials in sections S2.2 and S2.3. Temporal trends found in the longer citation windows are largely consistent with our findings in the main text.

Table 1. *Published papers, 1996–2014, and citations within two years of publication.*

Broad Category	Published Papers	Citations Within Two Years	Mean Citations Per Paper
Health	3,961,905	13,487,243	3.40
Social Sciences	851,893	1,410,878	1.66
Mathematics & Computer Sciences	1,174,541	1,644,994	1.40
Humanities	369,712	90,401	0.24

Compiled from the Web of Science (Clarivate Analytics). See Supplementary Materials, section S1, for the composition of the four broad categories listed above. One exceptionally highly cited paper in Math & CS is omitted.

of the total citations received by papers published in a given year; and the Herfindahl-Hirschman index (HHI).[8]

3 Monte Carlo evidence of bias in measures of citation inequality

Our core claim is that much of the apparent decrease in citation concentration is *not* the result of changes in the underlying pattern of inequality in citations, but instead an artifact of increases in the total papers published per field each year as well as growing numbers of total citations sent to those papers from subsequent publications. To demonstrate the theoretical plausibility of this claim, we perform a series of Monte Carlo simulations of four separate time series of hypothetical papers and incoming citations to those papers. In these experiments, we impose a counterfactual, fixed pattern of inequality while varying the total number of papers and citations based on the observed quantities from each field and year of the real-world data. Put simply, our simulations assume that the total number of papers and citations increases as in the real world, but that the distribution of citations follows a simple, fixed "power-law–like" pattern that does not vary over time or fields. If the Gini coefficient and other

8 HHI is a commonly used measure of market concentration computed by summing the squared market share of each firm. In our context, the market share is the citation count that one paper receives divided by the total citation count. Usually, when the HHI is smaller, it means the market is more decentralized; however, HHI also tends to decrease as the number of participants rises. For example, when 10 companies equally share a market, HHI is $0.1^2 \times 10 = 0.1$, but when 100 companies equally share, HHI is $0.01^2 \times 100$, or 0.01. As this illustration shows, *ceteris paribus*, HHI will decrease if total publication counts increase.

commonly employed measures of inequality were truly unaffected by marginals, they would find the same degree of citation concentration across these experiments. Instead, we show that inequality measures can be dramatically biased when comparing citation distributions with varying total papers and citations.

Formally, denote the ith paper published in field j and year t as p_{ijt}. Call the set of all such papers P_{jt}, which includes $|P_{jt}|$ total papers. Next, denote as n_{jtk} the number of future papers (over some chosen window of years) citing exactly k members of P_{jt}. (Observe that the sum of kn_{jtk} for all $k \in \{1, 2, 3, \ldots\}$ is also the total number of citations, N_{jt}, made to papers in P_{jt} over the chosen window.) In the following, we focus on the potential distortion of inequality measures caused by variation in the marginals of these paper and citation distributions across time and fields: specifically, variation in $|P_{jt}|$ and n_{jtk}.

Drawing on these marginal quantities from the observed distributions of papers and citations, but no other real-world information, our Monte Carlo simulation consists of the following four steps:

Step 1: Define the set of hypothetical citable papers published in year t. For each year t and field j, create a set of hypothetical papers $\phi_{ijt} \in \Phi_{jt}$, distinct from the empirically observed set of published papers $p_{ijt} \in P_{jt}$. Let $|P_{jt}| = |\Phi_{jt}|$, so that the number of hypothetical papers matches the empirically observed count for that field and year.

Step 2: Define the aggregate number of citations sent back to papers published in year t. Let ν_{jtk} indicate the number of hypothetical future papers citing exactly k papers in Φ_{jt}. Then set $\nu_{jtk} = n_{jtk}$, so that the total number of hypothetical citations matches the total citations actually received by papers published in year t and field j. (This degree of specificity is required because each future paper must send a discrete number of citations.)

Step 3: Define a time- and field-invariant pattern of inequality in the distribution of incoming citations. For simplicity, we assume that papers come in four ranked categories: superstar papers (the top 1% of papers published in a field-year), star papers (the next 9% of papers), solid papers (the next 20% of papers) and weak papers (the bottom 70%). When a paper sends an additional citation to a paper published in year t, we assume that citation is r times more likely to land on a given superstar paper than on a given star paper. Likewise, that citation is r times more likely to be sent to a particular star paper than to a particular standard paper, and r times more likely to cite a given

standard paper than a given weak paper. In our simulations, we set $r = 4$, which implies that when a future paper adds a citation to a paper in Φ_{jt}, it is 64 times more likely to send that citation to a particular superstar paper than to a particular weak paper.[9]

Step 4: Simulate citations to papers published in year t by papers published in later years. For each field j and year t, simulate a single hypothetical future paper's bibliography by sampling without replacement k papers from Φ_{jt} using the probabilities defined in Step 3. We repeat this exercise ν_{jtk} times for each $k \in \{1, 2, 3, \ldots\}$ to build up the complete set of citations to papers published in field j and year t. We then count the number of times each paper in Φ_{jt} has been cited to create a simulated citation distribution. Finally, we summarize this distribution using each of our measures of inequality. (Step 4 should be repeated several times, averaging each measure of inequality across runs. We found even 10 simulations was sufficient to reduce Monte Carlo error to negligible levels.)

The only thing that varies across simulations for different fields and years is the marginal number of papers and citations to papers; we have held constant the underlying structure of inequality in how likely a specific paper is to receive a citation. Therefore, if the Gini coefficient (for example) is truly unaffected by field-specific or year-over-year changes in the marginals, we should observe the same Gini coefficient regardless of which field and year of marginals we use in the simulation.

We illustrate the logic of our Monte Carlo experiments using the example of the social sciences (Figure 2). As in other broad disciplines, the number of papers published in the social sciences – and the number of citations sent to those papers – have generally increased each year, with a particularly rapid rise in the first decade of the twenty-first century (left panel of Figure 2). To demonstrate the logic of marginals bias, our Monte Carlo experiments simulate a set of papers published, and citations to these papers, over a period of years. The pattern of inequality for incoming citations to these papers is fixed across years, but the total number of papers published and citations sent is set to match exactly the marginal quantities observed in the social sciences (middle panel of Figure 2). If the Gini coefficient were immune to marginals bias, these results – marked *Simulated with Fixed Inequality* – would be a perfectly flat line. Instead, the rising

9 This is the only step in the simulation that requires arbitrary calibration to choose a "reasonable" pattern of inequality likely to correspond to the real-world. A small amount of calibration suggests that the four-tiered structure assumed above closely matches real world citation distributions across fields when $r = 4$.

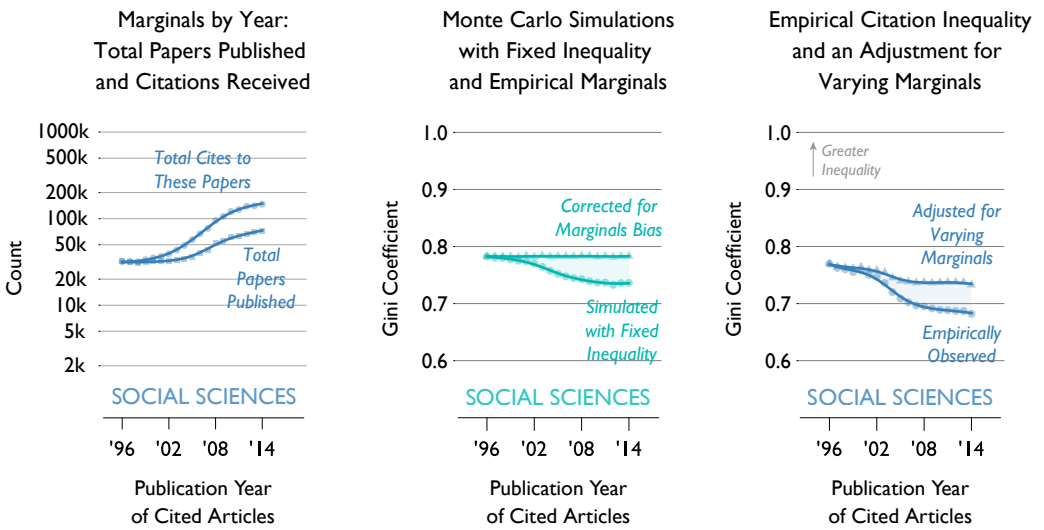

Figure 2. *How the social sciences illustrate the logic of marginals bias in citation counts.* The left panel shows empirical marginal counts of papers and citations within two years of publication over the years 1996–2014. The middle panel presents Monte Carlo simulations of the Gini coefficient for citations per paper, and a correction for marginals bias. The right panel reports empirical measures of the Gini coefficient for the social sciences, and the adjustment for changing marginals. All curves are smoothing splines with span of 0.5.

marginals of social science publications and citations produces a strong tendency to mistakenly infer declining citation inequality over time, even though the actual level of inequality in these simulations does not change. (As we shall see, this pattern also holds for other disciplines and even other inequality measures.) This result implies that Gini coefficients measured across years and fields with varying marginals are not directly comparable.

While the simulation results reflect an assumed pattern of citation inequality, it is worth noting how remarkably they resemble – both in terms of average levels by field and changes over time – the actual Gini coefficients obtained from the Web of Science data (shown in the right panel of Figure 2 as *Empirically Observed*), a pattern that will hold across disciplines and inequality measures. This suggests two hypotheses: first, that the "power-law-like" model of citations we adopt in our simulations is a plausible simple model of actual citation behavior; and second, that variation in total papers and total citations may have created the illusion of declining inequality over time when no such trend actually exists.

4 A resampling correction for bias in measures of citation inequality

Our Monte Carlo experiments suggest the Gini coefficient and other common inequality measures are unreliable guides for comparisons across time and fields, and should be avoided. However, if the "marginals bias" can be corrected, we think these tools can still be used. To do this, we introduce a resampling correction and an `R` package, `ineqReSample`, which allows users to correct inequality metrics computed on their own data.[10] The key idea behind our correction is to choose a base year, for which we observe the total number of papers published and the total number of citations to those papers that follow. For each subsequent year, we resample the papers published in that year and the citations to them to have the same marginals as observed in the base year – thus preserving the underlying time-varying structure of citation inequality but in samples drawn with fixed total numbers of papers and citations.

Inequality measures computed based on resampled citations should be comparable relative to the base year for each field. This suggests that our adjusted measures could be employed in an analogous fashion to other metrics that need adjustment to a base year, such as seasonality or inflation adjustments in economic research (though we emphasize the causes of marginals bias are distinct from the processes underlying inflation and seasonal variation in economic data).

In the simplest case, the number of papers and citations are at their minima in an initial reference year. Adjusting inequality measures in subsequent years to be comparable to the initial year involves four steps:

Step 1: Sample to match the original total number of papers. For each year $t > 1$, sample without replacement $|P_{j,1}|$ citable papers from P_{jt}; call this subsample of papers Q_{jt}.[11]

10 The `ineqReSample` package is available at `https://github.com/lanukim/ineqReSample`

11 The algorithm is somewhat more complex if the base year has more total publications than some other year. If for some year z, $|P_{jz}| < |P_{j,1}|$, then it will not be possible to sample without replacement a sufficient number of papers from the original pool of year z publications. Sampling with replacement is not a solution, as any papers that are sampled twice (or more) will appear to have double (or more) citations compared to their actual citation counts. A simple solution that preserves the original distribution of papers and citations is to create a set of "duplicate" papers P'_{jt} with the same pattern of citations as the observed papers P_{jt}. We then sample from the combined set of papers in $\{P_{jt}, P'_{jt}\}$ as if the actual pool of published papers were twice the observed size, but with the same empirical distribution of citations per paper. (If

Step 2: Sample incoming citations to match the original number of total citations. From all the cites to papers in Q_{jt}, sample without replacement $N_{j,1}$ citations.[12]

Step 3: Compute comparable measures of inequality using the sampled citations to the sampled papers. These might include Gini coefficients, percentage of papers ever cited, quantiles-based measures, the Herfindahl-Hirschman index, and other metrics.

Step 4. Repeat steps 1–3 and average the results to reduce sampling error. Even a small number of simulations is sufficient to reduce sampling error to negligible levels, though more should be used if the total number of papers and citations is low.

We demonstrate the accuracy of our resampling correction by first applying it to our simulation results, where we know the only potential explanation for varying Gini coefficients across time are changes in the total number of papers and citations. The line marked *Corrected for Marginals Bias* in the middle panel of Figure 2 shows that the resampling-corrected Gini detects no change in the level of inequality over time. Thus our Monte Carlo experiments show that this procedure successfully removes all of the bias introduced by changing marginals in the social sciences. (The same holds for each broad discipline and measure of inequality considered herein.) Our simulation-based adjustment has rendered the Gini coefficient comparable across years with varying marginals, revealing a common underlying pattern of inequality.

We next apply this adjustment to the empirical citation data for the social sciences (*Adjusted for Varying Marginals*, as shown in the right panel of Figure 2). We expect unadjusted Gini coefficients to be non-comparable due to rising marginal counts of papers

---

needed, this process can be repeated to make the pool of published papers as large as necessary without altering the behavior of the citation distribution.) This technique does not alter our results in cases where $|P_{jz}| \geq |P_{j,1}|$ and makes possible corrections to inequality measures when $|P_{jz}| < |P_{j,1}|$.

12 Adjustment to match the marginal citations from the first year is not feasible for any year whose papers have fewer incoming citations than the initial, reference year: it is not possible to sample $N_{j,1}$ citations from N_{jt} without replacement. In our data, this is only an issue for the two-year citation window for the some of the early years of Humanities (1997–2002) when total citations dipped slightly below 1996 levels. Alternative solutions in this case include choosing a different base year (the year with the fewest total citations) or adjusting all years to a total number of citations below the observed levels. Instead, to keep comparisons of our results across fields as straightforward as possible, we omit the adjusted results for Humanities for the years 1997–2002 from Figures 2–6. Because the marginals for Humanities were largely stable over this period, the omission has little effect on our findings.

and citations, with a bias towards reporting declining inequality even if there is little or no actual reduction in the concentration of citations. Our adjustment shows this concern is warranted: the large majority of the ostensible reduction in the Gini coefficient appears to be an artifact of increasing marginals. Adjusting for these varying marginals reveals only a small reduction in Gini overall, and essentially no change in citation inequality after 2005.

5 Adjusted measures of citation inequality by field and indicator

In the remainder of the paper, we report Monte Carlo results for each field and inequality measure and explore what happens when real-world citation data from each of the four broad disciplines are adjusted for marginals bias.

5.1 Gini coefficient

We now expand our Monte Carlo simulation of the Gini coefficient across fields as well as years. The lines marked *Sim* in the top half of Figure 3 show the Gini coefficient of the citation distribution in simulations that assume a fixed pattern of inequality over time and fields, but the same marginals as in the papers observed in Web of Science for that field and year. These simulations demonstrate that increasing marginals are sufficient to produce the illusion of declining year-to-year Gini coefficients, even if patterns of inequality remain constant. Moreover, for each field, the simulations track fairly closely with real world data (marked *Obs* in the lower half of Figure 3), suggesting that the real world increase in citations may be an artifact of changing marginals, and not an indication of greater diffusion of citations. Once the simulated Gini coefficients are adjusted for changing marginals, they show no change over time in any field (see the lines marked *Cor* in the top half of Figure 3). While the fields themselves still appear to have different levels of inequality after correcting for marginals, this is only because we have adjusted each time series of Gini coefficients to be comparable to the base year for that field. Creating inter-field comparable measures would require us to impose the same marginals to all fields in the resampling correction. [13]

13 We do not take this step in the initial sets of analyses, in order to focus on comparison between the observed and adjusted levels of inequality in each field. However, in Figure 8 and 9, we remove marginals bias that hinders inter-field comparison by adjusting marginals to have the same number of papers and citations to all fields except humanities. We exclude humanities

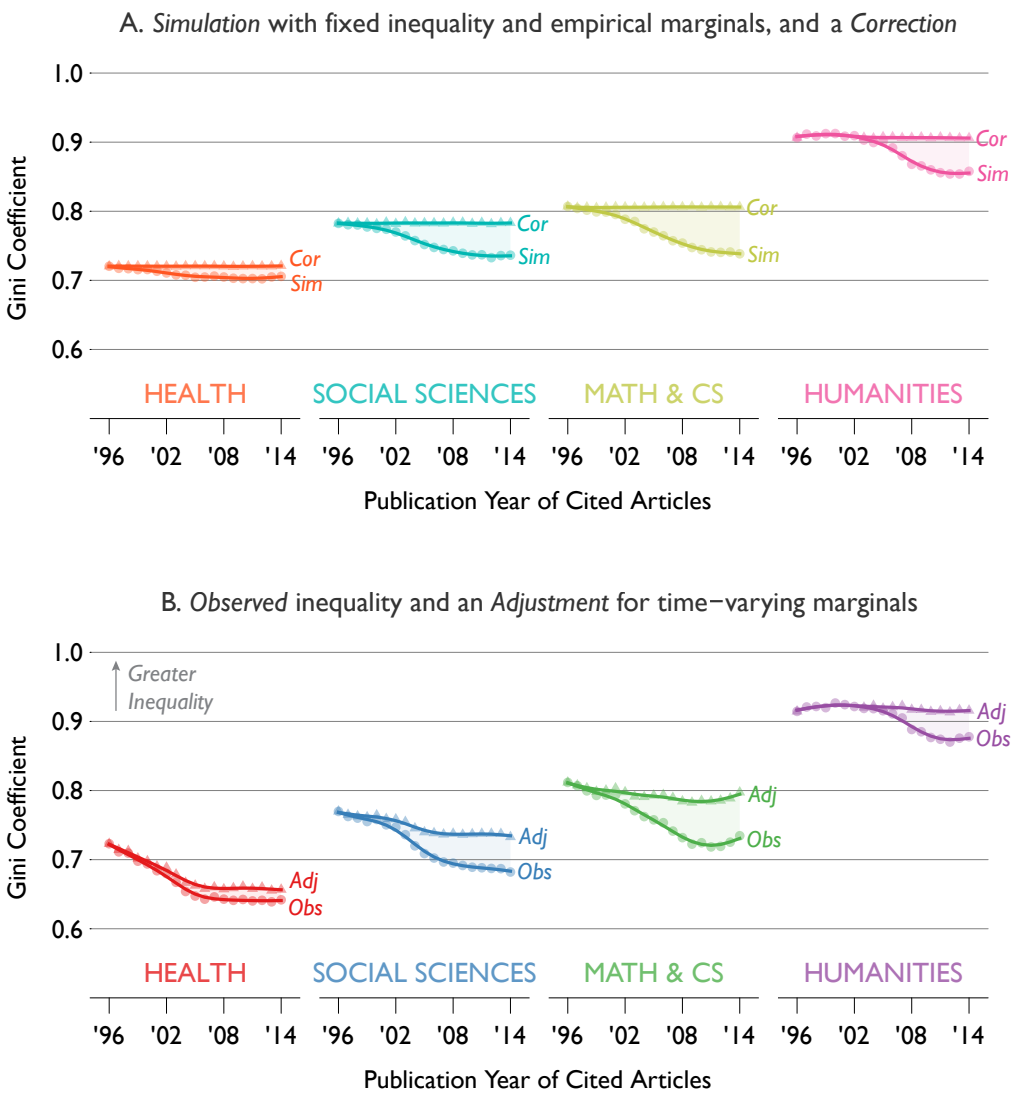

Figure 3. *Gini coefficient for citations within two years of publication, 1996–2014: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show Gini coefficients of citation distributions from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in the Gini coefficient using a resampling correction. The lines marked *Obs* in the bottom panel shows Gini coefficients over fields and time using the empirical data from Web of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2002. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

We now apply this approach to the actual empirical citation data for each field. The lower panel of Figure 3 shows two versions of the Gini coefficient calculated by field and year using the Web of Science data: an uncorrected version (marked *Obs*) potentially biased by changing marginals, and an adjusted version (marked *Adj*) that renders the Gini coefficients comparable (across years within the same field only) by resampling papers and citations to match the totals in the first year of each field's time series. Without adjustment, as in the prior literature, there appears to be a trend towards lower concentration of citations in most fields, with the greatest change in the first decade of the 21st century. However, adjusting for marginals reveals that this reduction in inequality is mostly a mirage. In the humanities, for example, the Gini coefficient appears to have changed not at all once the dramatic increase in citations over this period is accounted. Likewise, the Gini for the social sciences and for mathematics and computer sciences appear to have fallen only slightly, with the vast majority of the apparent decrease merely an artifact of growth in papers and citations. Only in health, where the number of papers and citations to papers were already very high in 1996, does the apparent decrease in concentration appear genuine, though it is worth noting that inequality in health publications appears to be essentially constant after 2005.

5.2 Proportion of ever cited papers

The percentage of papers ever cited is both the simplest measure of citation concentration and the measure most likely to be affected by marginals bias. The logic is straightforward: if any given paper has a fixed non-zero probability of being cited by each subsequent paper, the probability of having at least one citation will increase as the total number of future papers and citations increases.

Here, we examine whether the share of papers cited within two years of publication is subject to marginals bias using both Monte Carlo simulation and the Web of Science corpus. The unadjusted observed papers ever cited (marked *Obs* in the lower half of Figure 4) are quite similar to earlier estimates from Larivière et al. 2009 and show differences across broad disciplines in the percentage of papers ever cited as well as generally upward trends in papers ever cited (i.e., declining concentration in citations).[14]

---

from these comparisons because its smallest marginals – for the early years of the humanities – are so much lower than other fields as to make cross-field comparison particularly difficult.

14 When we compare our results to Larivière et al., we focus on the years 1996 to 2005 and the two fields (social sciences and humanities) that most closely mirror Larivière et al.'s analyses.

However, our Monte Carlo experiments reveal that the percentage ever cited is the inequality measure most affected by marginals bias. The lines marked *Sim* in the top half of Figure 4 show the percentage of papers receiving any citations in simulations that assume a fixed pattern of inequality over time and fields, but the same marginals as in the papers observed in WoS. The simulations not only provide an eerily close match with the real world data, they also show that increasing marginals are sufficient to produce rising percentages of papers ever cited, even if patterns of inequality remain constant. This suggests the real world increase in the percentage of papers receiving citations may be an artifact of changing marginals, and not an indication of greater diffusion and diversity of citations. When we adjust the observed percentage ever cited for marginals bias (marked as *Adj* in the lower half of Figure 4), the trend towards higher percentage virtually disappears, with the partial exception of publications in health.[15]

5.3 Proportion of papers accounting for 20% and 80% of citations

Whereas the unadjusted percentage-ever-cited metrics of inequality (unreliably) suggest in recent years declining citation concentration, unadjusted quantile based measures such as the the proportion of papers accounting for 20% and 80% of total citations made in a given year offer less evidence of declining inequality. The results marked *Obs* in the lower half of Figure 5 show that the percentage of papers accounting for 20% of citations was rising in the health and social science disciplines from 1996 to about 2007, matching Larivière et al's findings. But after that, this measure of concentration is flat, suggesting stable patterns of inequality. In mathematics and computer sciences, there may even be a shrinking percentage of papers accounting for 20% of citations after 2008. Only in the humanities do the *unadjusted* data suggest falling concentration after 2008.

But to what extent are these apparent trends affected by changing total papers published and cited? Because of the well-known robustness of quantile measures of distributions, we expect these metrics to be less affected by marginals bias. Moreover, to the extent marginals bias is driven by the papers at or near the lower zero bound of citations, we expect bias to be especially small for quantiles that mainly capture concentration at the top of the citation distribution, such as the percentage of papers accounting for 20% of all citations.

The top half of Figure 5 presents our Monte Carlo results, which suggest that the degree of marginals bias should be small for the broad disciplines of health, the social

15 The impact of increased publications and longer reference lists in newer publications on the proportion of ever cited papers has been also found by Wallace et al.2009.

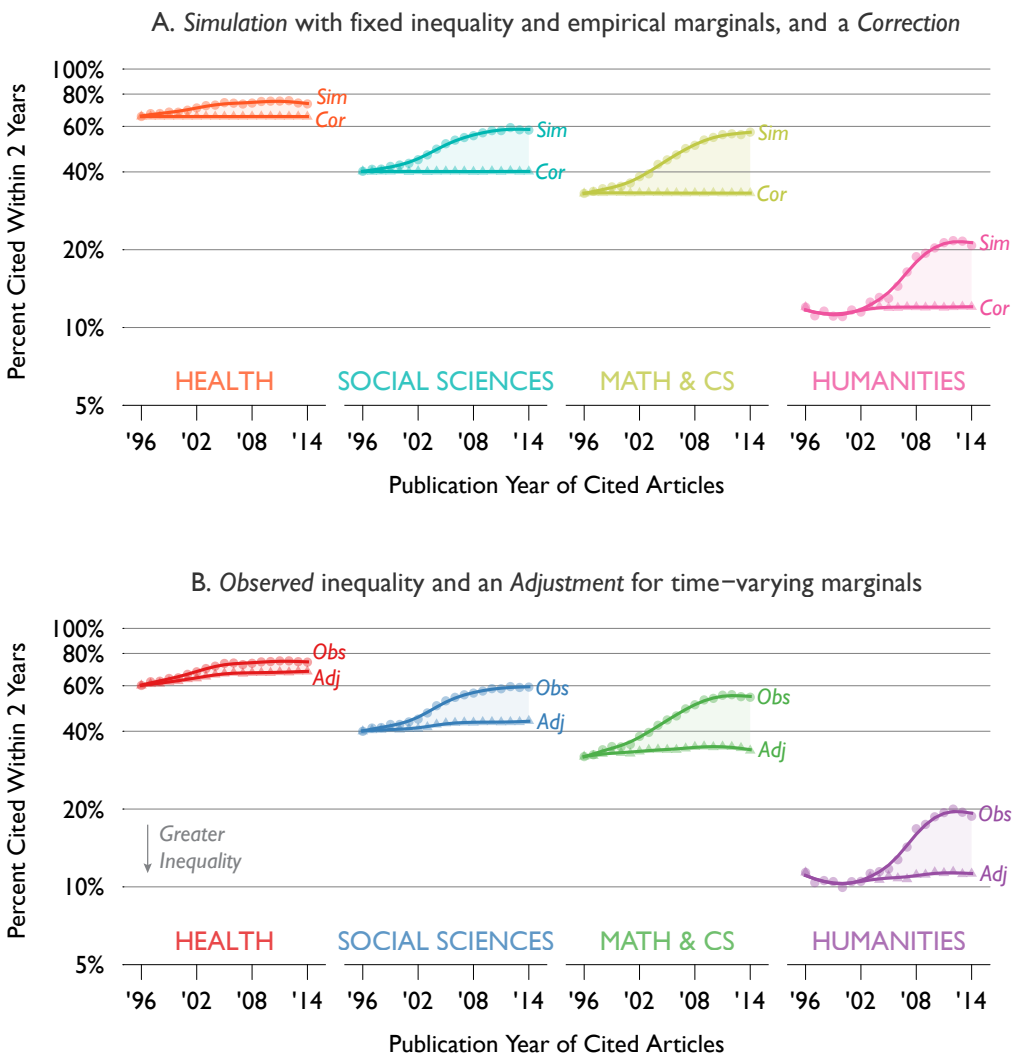

Figure 4. *Percent of papers with any citations two years after publication, 1996–2014: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers ever cited from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in percent-ever-cited using a resampling correction. The lines marked *Obs* in the bottom panel shows percent-ever-cited over fields and time using the empirical data from Web of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2002. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

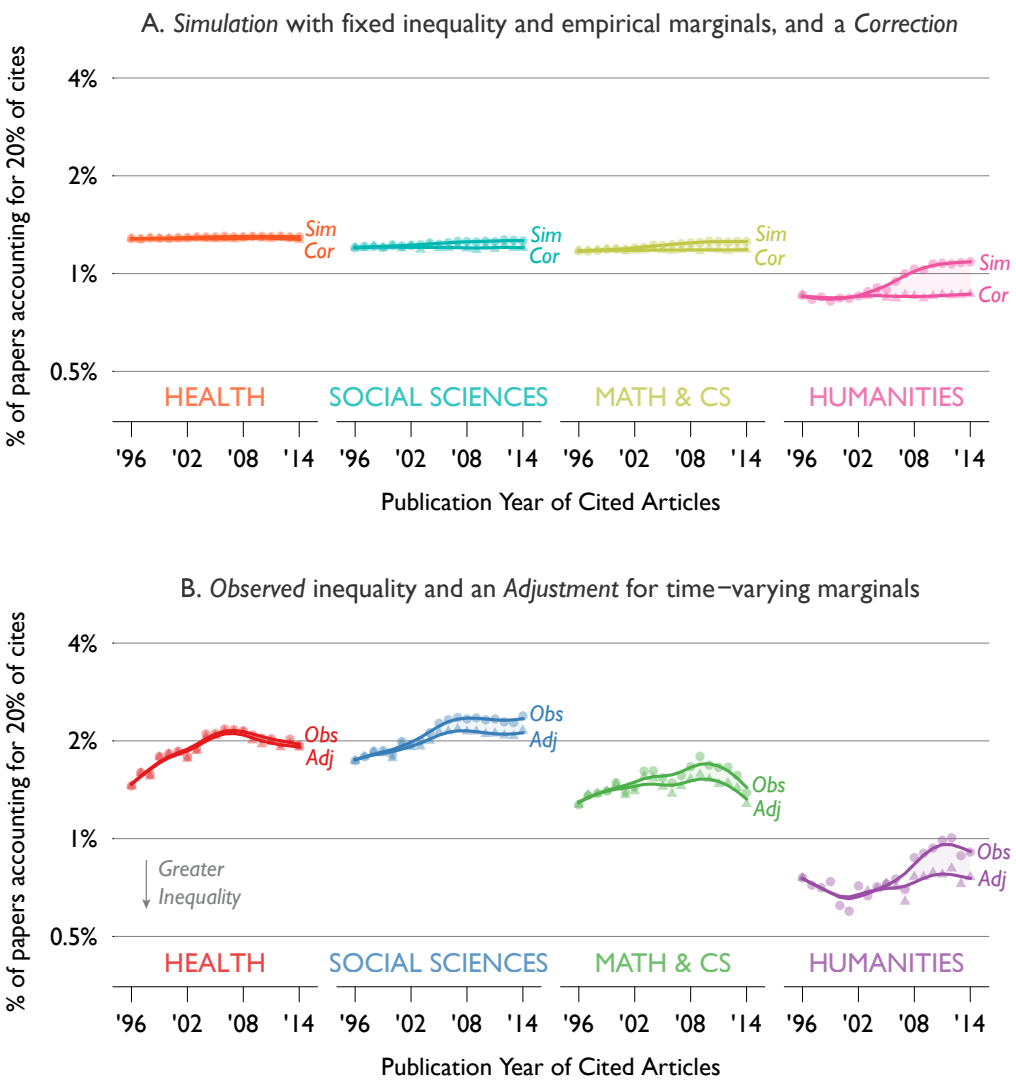

Figure 5. *Percent of papers accounting for 20% of all citations within two years of publication, 1996–2014: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers accounting for 20% of all citations from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias using a resampling correction. The lines marked *Obs* in the bottom panel shows the percent of papers accounting for 20% of citations over fields and time using the empirical data from Web of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2002. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

sciences, and mathematics and computer sciences. In these fields, total papers and citations are substantial enough – and the top 20% of citations likely concentrated enough – that the presence of varying numbers of papers near the zero bound is unlikely to substantially bias this metric. Humanities, on the other hand, appears to be subject to considerably bias even in quantile measures of inequality as a result of its small and rapidly shifting total citation count.

The bottom half of Figure 5 confirms these intuitions: the results for health, the social sciences, and mathematics and computer sciences are largely unaffected by adjustment. However, the appearance of growing equality in the humanities after 2008 proves to be an illusion: adjusting for margins, the percentage of humanities papers accounting for 20% of citations has barely shifted since 1996. Overall, then, once adjusted for margins, there is no evidence in any broad discipline for declining inequality in this metric in the most recent decade of available data.

Turning to our second quantile-based measure, the percentage of papers accounting for 80% of citations over a two-year window, we find a pattern more similar to that of the Gini coefficient. Our Monte Carlo results (top panel of Figure 6) suggest there may be substantial marginals bias in this measure for the social sciences, math and computational sciences, and humanities, with only health – with its much larger number of total papers and citations – largely immune. This fits the intuition that even quantile-based measures can suffer from marginals bias if they focus on parts of the citation distribution that are likely to be strongly influenced by the proportion of papers at or near the zero lower bound on citations.

Looking at the Web of Science corpus, we find the unadjusted percentage of papers receiving 80% of citations rises in all fields, though mostly in the earlier years of our data. However, adjusting for marginals eliminates virtually all of the reduction in inequality. Once the changing total number of papers and citations is accounted for, it appears once again that only citations to pre-2006 health papers show evidence of a trend to greater equality. In other fields, particularly the humanities and mathematics and computer sciences, the adjusted percentage of papers accruing 80% of citations is essentially unchanging over time.

5.4 Herfindahl-Hirschman Index

Finally, we apply the same analysis to the Herfindahl-Hirschman Index (HHI). Computing the unadjusted HHI from the observed data from Web of Science suggests declining concentration in all broad disciplines except health, where HHI is mostly con-

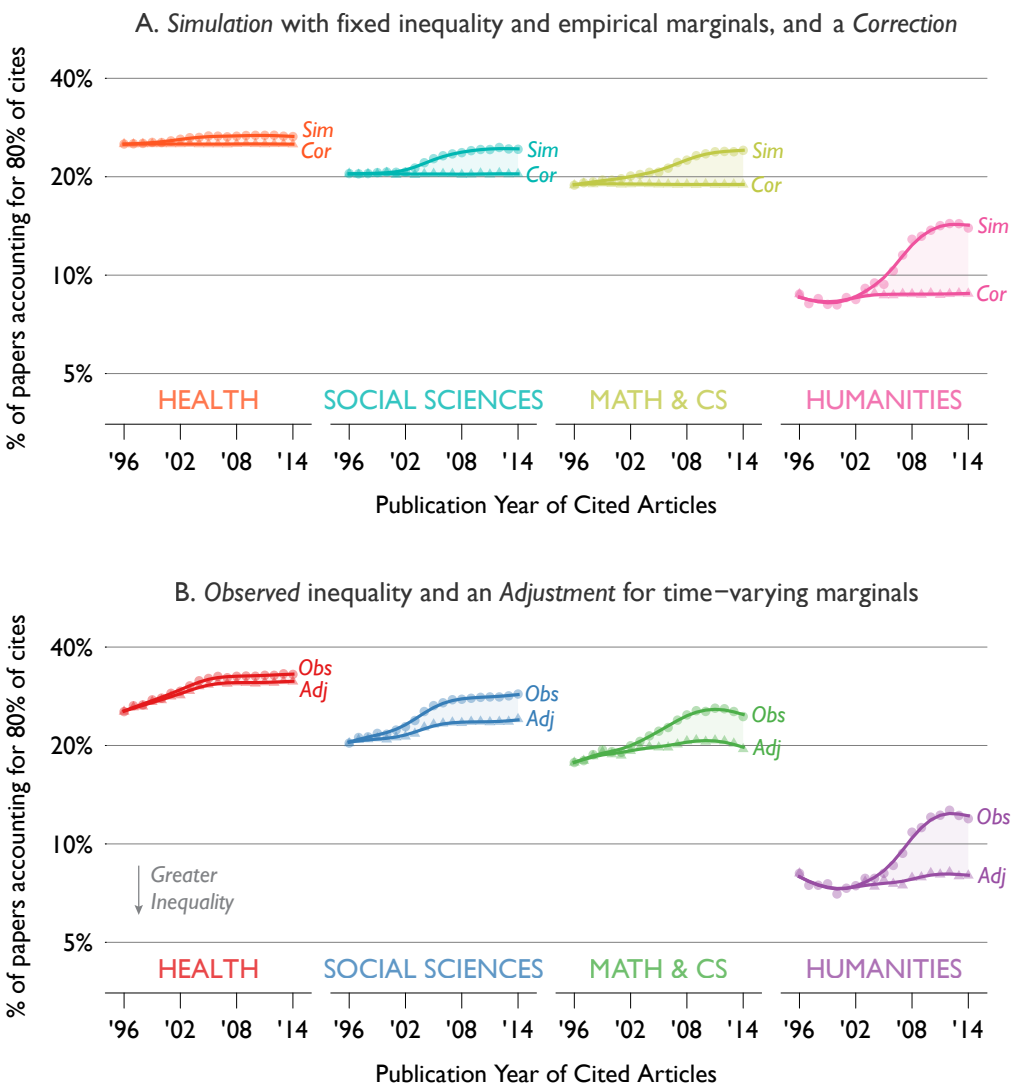

Figure 6. *Percent of papers accounting for 80% of all citations within two years of publication, 1996–2014: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers accounting for 80% of all citations from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias using a resampling correction. The lines marked *Obs* in the bottom panel shows the percent of papers accounting for 80% of citations over fields and time using the empirical data from Web of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2002. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

stant with a slight increase since 2008 (see lines marked *Obs* in the lower half of Figure 7), matching the findings of Larivière et al. However, our Monte Carlo experiments suggests HHI for all four fields may be subject to a substantial degree of marginals bias (see the lines marked *Sim* in the top half of Figure 7). Applying our adjustment to HHI for the observed data reveals all of the apparent reduction in concentration to be an artifact of increasing total publications and citations over time. The adjusted HHI is essentially constant over time for the humanities, the social sciences, and the mathematics and computer sciences. And in health, we find evidence that inequality has actually increased since 2007, once changing marginals are taken into account.

6 Adjustment to fixed marginals across fields and time

In the preceding section, for each publication year after 1996, we resampled papers and citations to have the same totals as in 1996 *by field*. This strategy allowed us to trace within-field changes in citation inequality without being misled by marginals bias. We can also accurately note whether inequality is changing in similar ways across fields. In short, adjusting each field to its own set of references margins allowed us to address our primary research questions. However, *inter-field* comparisons of the average level of inequality predominant in each field are still susceptible to marginals bias unless we adjust the total papers and citations to a common set of margins across fields. In other words, if we wish to assess which fields tend to be more concentrated or diffuse in their citations on average across time, we will need to make further adjustments for varying marginals across fields.

To allow such inter-field comparisons for health, social sciences, and mathematics and the computer sciences, the results reported in this section resample each field-year of published papers and citations to those papers to have the same total counts (30,000 papers and 30,000 citations) regardless of field or year.[16] We refer to metrics computed from these marginals as "fully adjusted." We exclude the humanities (which had far fewer papers and citations, especially in the earlier years) from the fully adjusted

16 The choice to set *both* papers and citations to the same number – 30,000 – is a coincidence driven by the minima of the observed distributions of papers and citations across these fields and years. It would be perfectly reasonable to set the total number of papers to a different common marginal than the total number of citations, so long as each marginal was kept the same across fields and years.

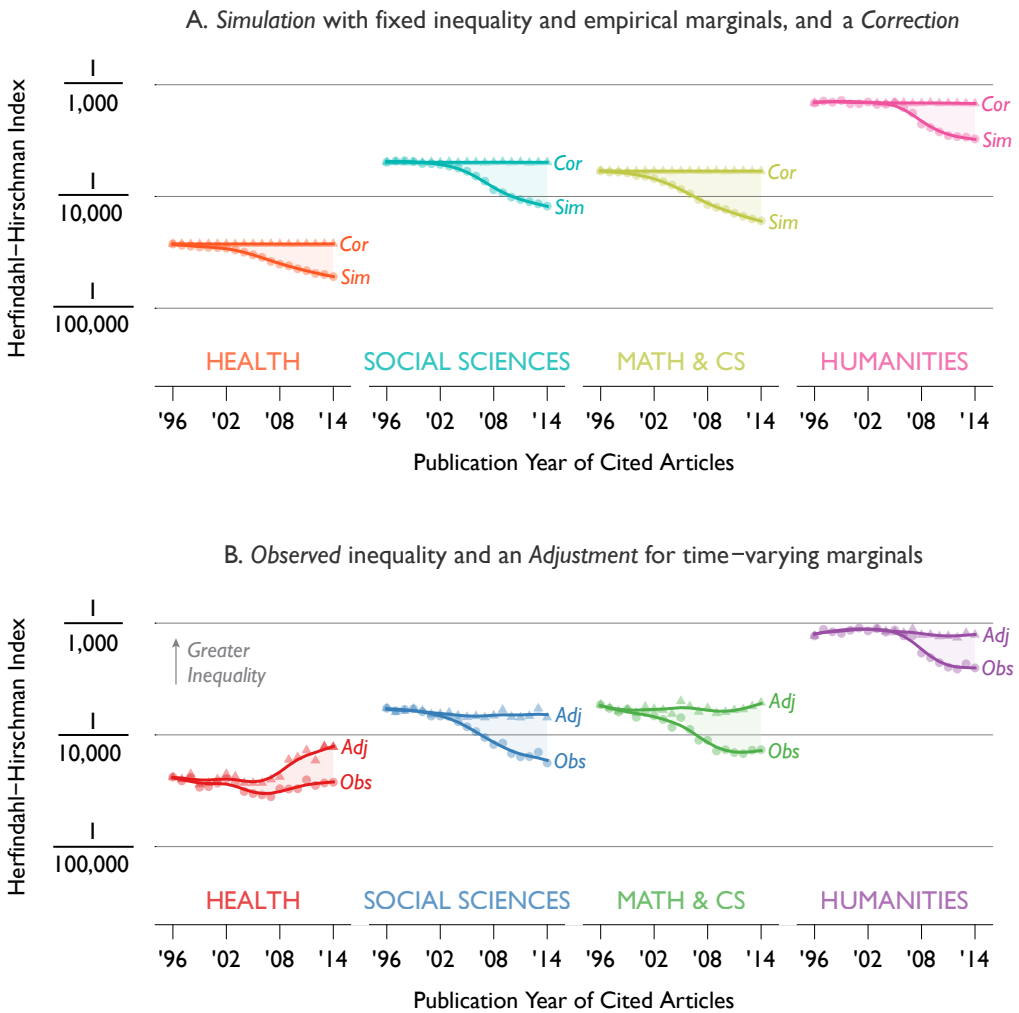

Figure 7. *Herfindahl-Hirschman Index of citations within two years of publication, 1996–2014: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show the Herfindahl-Hirschman Index (HHI) of citation concentration from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in HHI using a resampling correction. The lines marked *Obs* in the bottom panel shows HHI over fields and time using the empirical data from Web of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2002. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

comparison to avoid using uncomfortably small marginals, particularly for citations. Throughout this section, we use the same 2-year citation window.

In Figures 8 and 9, we report all five metrics of inequality under fully adjusted marginals. Overall, full adjustment reveals that most inter-field differences in inequality levels are due to different marginals between fields. For example, the results reported in Figures 3, 4, 6, and 7 suggested that on most metrics, the health field seemed to have less inequality overall than other fields when margins are adjusted to field-specific reference years. However, this apparent difference is just another example of marginals bias. After we resample all three broad fields to have the same marginals, health and the social sciences have similar levels of concentration and similar trends when measured by the Gini coefficient (Figure 8), the percentage of papers ever cited (Figure 8), and the percentage of papers accounting for 80% of citations (Figure 9). On the same three metrics, we find that citation concentration in math and computer sciences is slightly higher than other two broad fields regardless of year. However, comparing the fully adjusted Herfindahl-Hirschman index (Figure 8) and percentage of papers accounting for 20% of citations (Figure 9), we find inequality in health may even be slightly higher than in the social sciences, while mathematics and computer sciences appear more similar to health. These differences across metrics likely reflect concentration at different points in the distribution. As HHI and the percent of papers accounting for 20% of citations are more sensitive to concentration at the top of the distribution than our other metrics, we infer that citations in mathematics and computer sciences as well as health may be slightly more concentrated at the top of the distribution than citations in the social sciences. Looking across the whole distribution, math and computer sciences may be somewhat more concentrated than either health or the social sciences.

Finally, we see hints that citation concentration at the top of the distribution (as shown by HHI and the percentage of papers accounting for 20% of citations) is rising in recent years in mathematics and the computer sciences. However, all of these differences are very small: the key finding is that citation inequality is very similar not only over time, but across fields as well. Thus the results from inter-field comparison suggest that full adjustment for varying marginals is essential for meaningful comparison of citation concentration across fields.

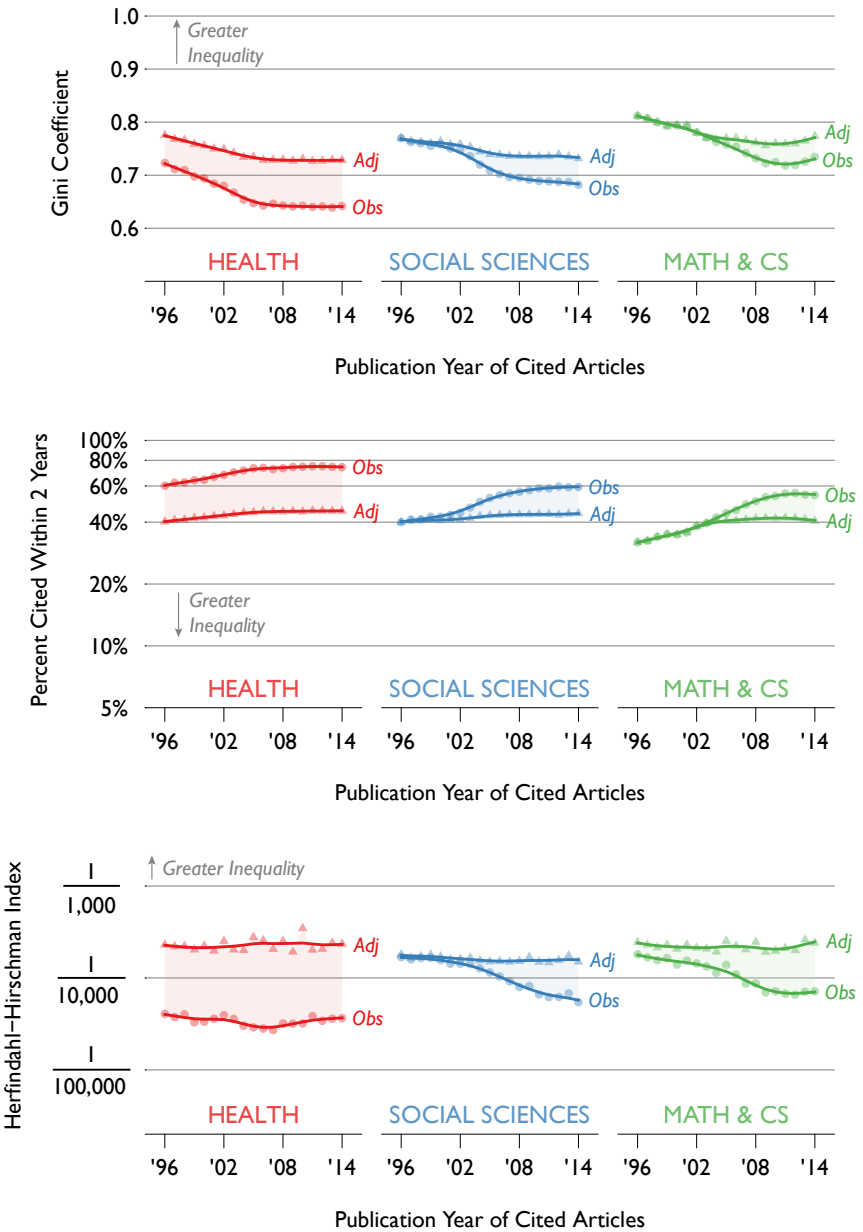

Figure 8. *Gini coefficient, percent of papers with any citations, and Herfindahl-Hirschman Index for citations within two years of publication, 1996–2014: empirical results with adjustment to fixed margins across fields.* All lines report results using empirical data from Web of Science. Lines marked *Obs* are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to a total of 30,000 papers published per year and 30,000 citations sent back to those papers over the following two years, regardless of field. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

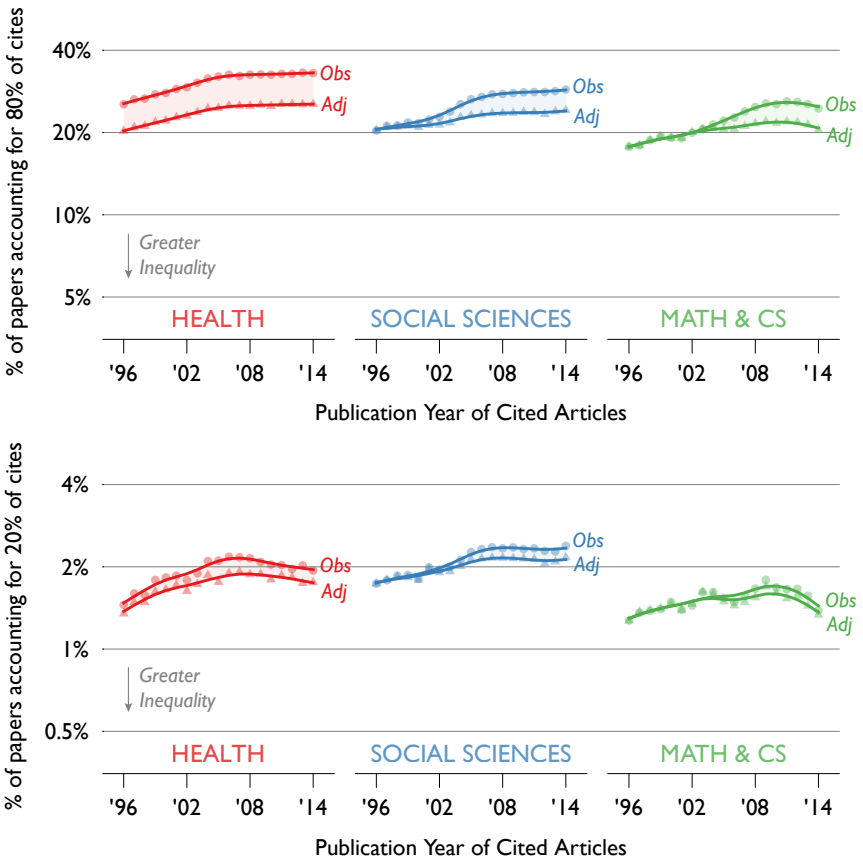

Figure 9. *Percent of papers accounting for 80% and 20% of all citations within two years of publication, 1996–2014: empirical results with adjustment to fixed margins across fields.* All lines report results using empirical data from Web of Science. Lines marked *Obs* are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to a total of 30,000 papers published per year and 30,000 citations sent back to those papers over the following two years, regardless of field. All curves are smoothing splines with span of 0.5. One outlier in Math & CS is omitted.

7 Discussion and conclusion

In this study, we identify the existence of marginals bias that affects inequality measures used to study scholarly citations. We then propose a resampling correction method that removes the bias. After adjusting measures of inequality to account for increasing marginals, we find minimal over-time change in the distribution of citations in most fields. Moreover, when we fully adjust marginals to give all fields the same number of papers and citations, there is little inter-field difference in citation inequality. This substantive finding is revealed only after adjusting for the substantial changes in the number of papers published and citations made during the period we study. Failing to adjust for these changing marginals when using a variety of metrics – including the Gini coefficient, percentage of papers with any citation, various quantile measures, and HHI – has lead some previous authors to conclude that there has been a decrease in the level of inequality in citations, and that scientific attention has become more diffuse (Larivière et al., 2009; Huang et al., 2012; Ranasinghe et al., 2015; Yoon et al., 2017). We believe this conclusion is incorrect, as are many of the conclusions based on comparisons of inequality across time and between groups (Evans, 2008; Diem & Wolter, 2013; Varga, 2019). Moreover, we suspect marginals bias may affect other inequality measures not directly addressed in this article. For example, a small amount of Monte Carlo experimentation suggests the Theil index is also subject to substantial marginals bias, which our adjustment appears to correct.

Monte Carlo experiments presented in this article and its Supplementary Materials suggest that while increases in the number of publications and citations lead to downward bias in inequality measures, the magnitude of the marginals bias effect varies. What explains this variation? We believe the most likely explanation is the coarseness of discrete measures, especially near the zero lower-bound for citations. As the total number of papers and citations rises, a smaller proportion of papers are likely to fall at or near the zero lower-bound, and citation counts in general are likely to be more informative. This fits with the smaller downward bias that appears when the marginals from the health field are used in simulations: the health field in general had the greatest number of papers and citations, and the smallest proportion of uncited papers. Similarly, the 6-year citation window, which accumulates more citations and reduces the share of papers receiving zero citation, is less vulnerable to this bias than the 2-year window. This logic also suggests that measures of inequality that are more sensitive to the extent of uncited or rarely cited papers – most obviously the percentage-ever-cited, but also Gini and HHI – will be more affected by varying marginals. In contrast, more robust

measures of inequality based on quantiles – such as the percentage of papers receiving *m*% of citations – should be less sensitive, particularly when they measure regions of the distribution that contain papers far from the zero lower-bound of citations.

Our results comparing adjusted inequality measures again highlight the fact that different measures of concentration and inequality capture different aspects of distributions (Piketty, 2014). For example, while it is empirically rare, it is theoretically possible for a distribution to be both highly concentrated and have a long tail. This is in fact what we observe in the Health field. As measured by HHI and the percentage of papers needed to account for 20% of citations, inequality in health citations has increased since the mid-2000s. Yet over the same period, the percentage of health papers ever cited and the Gini coefficient for health citations show a weak pattern of falling concentration. These differences between inequality measures imply that concentrated scientific attention on a small number of very highly cited papers may go hand in hand with a longer tail in the citation distribution. Thus, even after adjusting for marginals bias, scholars should carefully select inequality measures depending on what aspect of inequality is of most interest, or consider using a variety of measures to capture subtle differences in the pattern of concentration. For example, if concentration of citations to a very few highly cited papers is suspected, HHI or the percentage of papers needed to account for 20% citations (or an even smaller percentage) may be helpful. However, if the purpose of analysis is to measure a long-tail, either the proportion of ever cited papers or the percentage of papers needed to account for 80% of citations (or some other large percentage) would be most effective. The Gini coefficient essentially averages these tendencies, and therefore is less useful for investigating the specific nature of inequality.

Our conclusion challenges previous studies claiming that the scope of science has either narrowed (Evans, 2008) or broadened (Larivière et al., 2009). Instead, we found that the level of concentration in citation inequality has remained relatively stable. On the one hand, this stability could reflect a lack of fundamental change. While that would be consistent with our results, it is not the only possible explanation consistent with the evidence. If citation inequality is the product of several components, it could also be the case that stability is the result of well-balanced opposing forces. We consider two candidate forces: one social, and the other technological.

First, although we identify a method that effectively adjusts for the growth of publication and citation counts, we recognize that the increased volume of scientific papers itself is the result of important changes in the incentives, norms, and practices concerning the production and consumption of science. From the perspective of a producer,

the current generation of young scientists is under greater pressure to publish and be cited than prior generations (Warren, 2019) and an over-reliance on production metrics (Fire & Guestrin, 2019). From the perspective of a consumer of knowledge, scholars must adapt to the environment by allocating their limited time and energy to digesting the ever-growing volume of prior research (Pan et al., 2018; Parolo et al., 2015). Ultimately, the rising pressure to publish could result in an increase in the fraction of low impact publications, a social force that could lead to greater concentration in scholarly citations[17]

Second, there have been dramatic changes in the digital environments in which scholars search, read, and organize literature – in particular, technological innovations which, in principle, make it easier for researchers to keep up with a growing literature without devoting more time and effort to the task. If true, this could result in them citing a broader set of papers. Thus one possible explanation for the lack of change in the level of inequality in citation distribution is that scientists are using technological change to compensate for social change in the production of scientific papers. But even if this is the case, and the currently stable level of inequality is based on a balance of opposite effects, nothing guarantees these forces will remain balanced – especially if tighter academic labor markets accelerate scientific publication rates in the coming years.

However, it is also possible to speculate that technology might encourage greater concentration in scholarly attention in response to increasing pressure to publish, particularly in fields that move quickly such as computer sciences. Fast-moving fields frequently involve mass production of research results or strict conference deadlines, either of which may limit scholars' ability to read broadly. Our inter-field analysis in Figure 8 and 9 supports this conjecture by revealing that the math and computer sciences field has a slightly higher level of inequality than health and social sciences. The analysis also shows that the increase in concentration of citations toward the top of the citation distribution began around 2008, suggesting that computer scientists' early adoption of digital search tools , in combination with field-specific deadline pressures,

17 One could imagine the opposite direction as well. As more papers are published and as more subcommunities form in the literature, there may be a decrease in citation concentration. However, we think the argument for greater concentration is more plausible, given the likelihood of the Matthew effect in science (Merton, 1968). We encourage future research to sort these alternative hypotheses out, taking care to adjust inequality measures for changing marginals.

may have contributed to the concentration of academic interest towards a narrow set of highly cited papers.

While the empirical context for this study concerns scholarly citations, the methodological problem we identify extends to any context in which inequality measures are applied to indivisible count distributions containing many zeros. This pattern occurs when gatekeepers distribute scarce rewards across a large population; for instance, in the awarding of grants to investigators, and offers of admissions or jobs to candidates. In these examples, there are so few rewards per subject that comparison of inequality measures are vulnerable to the biases we identify in this paper. In a similar vein, we expect to find evidence of this bias in rapidly expanding markets for songs, movies, or books, especially if the volume of consumption is relatively stable. As we demonstrate, adjustment is particularly important in contexts where the target of behavior is discrete (as in citations or purchases) and many targets are rarely or never selected. To facilitate use of this method, we have created an open source `R` package, `ineqReSamp`, that adjusts inequality measures with the resampling correction. More details on the package can be found at `https://github.com/lanukim/ineqReSample`.

Of particular interest for future research is the impact that information retrieval technology (e.g., search engines and recommender systems) is having on what is found, read and cited in the scientific literature. Is technology narrowing or expanding our collective view of the literature? And what impact is this having on collective sense-making and ultimately to the success of science? In order to address these questions and related policy questions, we need measures that are unbiased, comparable over time and across fields, and reliably interpretable. We hope that our results revealing and correcting marginals bias will help advance research around these important questions.

References

Althouse, B. M., West, J. D., Bergstrom, C. T., & Bergstrom, T. (2009). Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology*, *60*(1), 27–34.

Atkinson, A. B. (1975). *The economics of inequality*. Oxford: Clarendon Press.

Barabási, A.-L., Song, C., & Wang, D. (2012). Handful of papers dominates citation. *Nature*, *491*(7422), 40–41.

Bornmann, L. & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222.

Brynjolfsson, E., Hu, Y., & Simester, D. (2011). Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science*, *57*(8), 1373–1386.

de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, *149*, 510–515.

Deltas, G. (2003). The small-sample bias of the Gini coefficient: Results and implications for empirical research. *Review of Economics and Statistics*, *85*(1), 226–234.

Diem, A. & Wolter, S. C. (2013). The use of bibliometrics to measure research performance in education sciences. *Research in Higher Education*, *54*(1), 86–114.

Evans, J. A. (2008). Electronic publication and the narrowing of science and scholarship. *Science*, *321*(5887), 395–399.

Evans, J. A. & Reimer, J. (2009). Open access and global participation in science. *Science*, *323*(5917), 1025–1025.

Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS biology*, *4*(5), e157.

Fire, M. & Guestrin, C. (2019). Over-optimization of academic publishing metrics: observing Goodhart's law in action. *GigaScience*, *8*(6), giz053.

Fleder, D. & Hosanagar, K. (2009). Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science*, *55*(5), 697–712.

Hamilton, D. P. (1990). Publishing by–and for?–the numbers. *Science*, *250*(4986), 1331–1332.

Hofstra, B., Kulkarni, V. V., Galvez, S. M.-N., He, B., Jurafsky, D., & McFarland, D. A. (2020). The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, *117*(17), 9284–9291.

Huang, M.-H., Chang, H.-W., & Chen, D.-Z. (2012). The trend of concentration in scientific research and technological innovation: A reduction of the predominant role of the us in world research & technology. *Journal of Informetrics*, *6*(4), 457–468.

Larivière, V., Gingras, Y., & Archambault, É. (2009). The decline in the concentration of citations, 1900–2007. *Journal of the Association for Information Science and Technology*, *60*(4), 858–862.

Larivière, V., Archambault, É., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and Technology*, *59*(2), 288–296.

Larsen, K. R., Hovorka, D., Dennis, A., & West, J. D. (2019). Understanding the elephant: The discourse approach to boundary identification and corpus construction for theory review articles. *Journal of the Association for Information Systems*, *20*(7), 15.

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63.

Pan, R. K., Petersen, A. M., Pammolli, F., & Fortunato, S. (2018). The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics*, *12*(3), 656–678.

Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., & Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics*, *9*(4), 734–745.

Petersen, A. M., Pan, R. K., Pammolli, F., & Fortunato, S. (2019). Methods to account for citation inflation in research evaluation. *Research Policy*, *48*(7), 1855–1865.

Piketty, T. (2014). *Capital in the 21st Century*. Cambridge, MA: Harvard University Press.

Ranasinghe, I., Shojaee, A., Bikdeli, B., Gupta, A., Chen, R., Ross, J. S., Masoudi, F., Spertus, J. A., Nallamothu, B. K., & Krumholz, H. M. (2015). Poorly cited articles in peer-reviewed cardiovascular journals from 1997-2007: Analysis of 5-year citation rates. *Circulation*, *131*(20), 1755–1762.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*(5762), 854–856.

Varga, A. (2019). Shorter distances between papers over time are due to more cross-field references and increased citation rate to higher-impact papers. *Proceedings of the National Academy of Sciences*, *116*(44), 22094–22099.

Wallace, M. L., Larivière, V., & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics*, *3*(4), 296–303.

Warren, J. R. (2019). How much do you have to publish to get a job in a top sociology department? or to get tenure? trends over a generation. *Sociological Science*, *6*, 172–196.

Weatherall, J. & O'Connor, C. (2020). Conformity in scientific networks. *Synthese*.

Yoon, S. J., Yoon, D. Y., Lee, H. J., Baek, S., Lim, K. J., Seo, Y. L., & Yun, E. J. (2017). Distribution of citations received by scientific papers published in the imaging literature from 2001 to 2010: Decreasing inequality and polarization. *American Journal of Roentgenology*, *209*(2), 248–254.

Zentner, A., Smith, M., & Kaya, C. (2013). How video rental patterns change as consumers move online. *Management Science*, *59*(11), 2622–2634.

Zollman, K. J. (2010). The epistemic benefit of transient diversity. *Erkenntnis*, *72*(1), 17.

SUPPLEMENTARY MATERIALS

To supplement Lanu Kim, Christopher Adolph, Jevin West, and Katherine Stovel, "The Influence of Changing Marginals on Measures of Inequality in Scholarly Citations: Evidence of Bias and a Resampling Correction," forthcoming in Sociological Science.

S 1 Aggregation of journals to disciplines and disciplines to fields

Across the Web of Science (WoS), journals are classified into one or more disciplines. If either of the first two disciplines listed fell into one of our broad categories, we include the journal in that field. We categorize broad disciplinary fields following National Science Foundation's taxonomy of disciplines created by the Integrated Postsecondary Education Data System survey. Out of 14 available categories, we use four broad categories that are more or less comparable to the analysis of Larivière et al. (2009).

S 1.1 Health

Allergy; Andrology; Anesthesiology; Audiology & Speech-Language Pathology; Cardiac & Cardiovascular Systems; Clinical Neurology; Critical Care Medicine; Dentistry, Oral Surgery & Medicine; Dermatology; Emergency Medicine; Endocrinology & Metabolism; Gastroenterology & Hepatology; Geriatrics & Gerontology; Health Care Sciences & Services; Health Policy & Services; Hematology; Infectious Diseases; Integrative & Complementary Medicine; Medical Ethics; Medicine, General & Internal; Medicine, Legal; Medicine, Research & Experimental; Neuroimaging; Nursing; Obstetrics & Gynecology; Oncology; Ophthalmology; Orthopedics; Pathology; Pediatrics; Peripheral Vascular Disease; Primary Health Care; Psychiatry; Public, Environmental & Occupational Health; Radiology, Nuclear Medicine & Medical Imagin; Radiology, Nuclear Medicine & Medical Imaging; Respiratory System; Rheumatology; Transplantation; Tropical Medicine; Urology & Nephrology; Veterinary Sciences

S 1.2 Humanities

Art; Classics; Dance; Ethics; Film, Radio, Television; Folklore; History; Humanities, Multidisciplinary; Literary Reviews; Literary Theory & Criticism; Literature; Literature, African, Australian, Canadian; Literature, American; Literature, British Isles; Literature, German, Dutch, Scandinavian; Literature, Romance; Literature, Slavic; Logic; Medieval & Renaissance Studies; Music; Philosophy; Poetry; Religion; Theater

S 1.3 Mathematics and computer sciences

Computer Science, Artificial Intelligence; Computer Science, Cybernetics; Computer Science, Hardware & Architecture; Computer Science, Information Systems; Computer Science, Interdisciplinary Applications; Computer Science, Software Engineering; Computer Science, Theory & Methods; Information Science & Library Science; Mathematical & Computational Biology; Mathematics; Mathematics, Applied; Mathematics, Interdisciplinary Applications; Statistics & Probability

S 1.4 Social sciences

Agricultural Economics & Policy; Anthropology; Archaeology; Area Studies; Asian Studies; Behavioral Sciences; Criminology & Penology; Cultural Studies; Demography; Economics; Ethnic Studies; Family Studies; Geography; Geography, Physical; Gerontology; History & Philosophy Of Science; History Of Social Sciences; International Relations; Language & Linguistics; Linguistics; Political Science; Public Administration; Social Issues; Social Sciences, Biomedical; Social Sciences, Interdisciplinary; Social Sciences, Mathematical Methods; Social Work; Sociology; Urban Studies; Women's Studies

S 2 Robustness checks

S 2.1 Analysis with a fixed set of journals

In this section, we test the robustness of the analysis using different restrictions on the data. First, in order to assess whether changes in coverage of journals in the WoS

database influenced the observed time trends, we limit the analysis to journals that published at least one paper continuously between 1996 and 2014 and were included in the WoS database during this entire period. Figure S1 shows the total number of papers published in the restricted set of journals by year and field, and the total number of citations to those papers from that restricted set of journals in the following two years. It is worth noting that small declines in total citations are more common using a restricted set of journals, which means we cannot perform corrections for marginals bias in the health field from 2013–2014, in the social sciences from 1997–1998, or in the humanities from 1997–2007.

When we repeat our analyses of citation concentration on this restricted dataset, our results are qualitatively similar to what we report in the main text. Nevertheless, there are some noteworthy new findings. First we review the results that hold in common for the Gini coefficient, the percentage of papers ever cited, the percentage of papers accounting for 20% and 80% of citations, and the Herfindahl-Hirschman Index (HHI). Looking at the unadjusted observed levels of citation inequality (marked *Obs* in the lower half of Figures S2–S6), we note the ostensible trend in the humanities towards greater equality has disappeared, suggesting that it was a function of new journals in conjunction with increasing total publications and citations. Focusing on only continuously published journals also reveals a slight uptick in the concentration in the broad area of health across all inequality metrics, whether adjustment is applied or not. Across all measures, the social sciences and mathematics and computational sciences remain strongly affected by marginals bias: although the unadjusted data suggests declining citation concentrations for these fields on each metric, the adjusted results (marked *Adj*) show that the inequality of citations has not changed for either discpline once marginals bias is removed. On balance, then, focusing on continuously published journals suggests no tendency towards greater equality in any field, and perhaps a movement in the opposite direction for health publications.

Finally, we note that all metrics still seem affected by marginals bias to the extent noted in the main text, with the partial exception of HHI, which is somewhat less biased – but still unreliable without adjustment – when a fixed set of journals is compared over time.

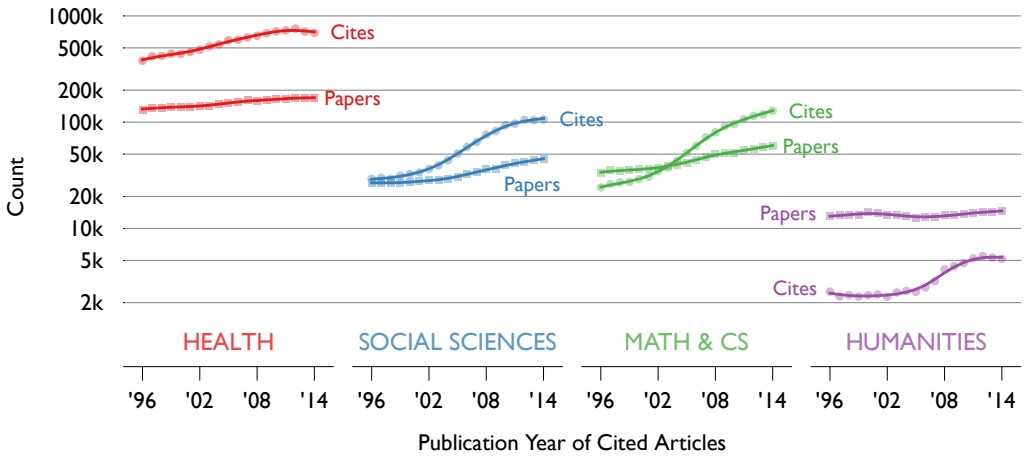

Figure S1. *Number of journal articles published 1996–2014 and citations to those articles within two years of publication with the fixed set of journals continuously published between 1996–2014.* Compiled from the Web of Science (Clarivate Analytics). Trend lines estimated by robust-and-resistant regression to minimize the influence of outliers. All curves are smoothing splines with span of 0.5.

S 2.2 Analysis with longer time windows: citations over four years

In order to assess the sensitivity of our results to the use of a relatively short two-year citation window, we repeat our analysis using four-year citation windows. Because our analysis includes citations made between 1996 and 2016, the four-year citation window only includes papers published up to 2012 (four years before 2016.) Figure S7 shows the total number of papers published in each of these years, by field, and the total number of citations to those papers in the following four years. Total citations increased in every field except the humanities from 1997–2000, which are the only cases that could not be adjusted for marginals bias.

The changes in citation concentration over time observed follow patterns similar to those produced using the shorter window in analyses presented in the main text (Figures S8–S12). The degree of marginals bias grows (almost imperceptibly) smaller as the citation window grows longer – and fewer published papers thus remain close to the zero-lower bound for citations – but does not disappear, suggesting adjustment is still necessary for longer windows. Part of the apparent reduction of marginals bias

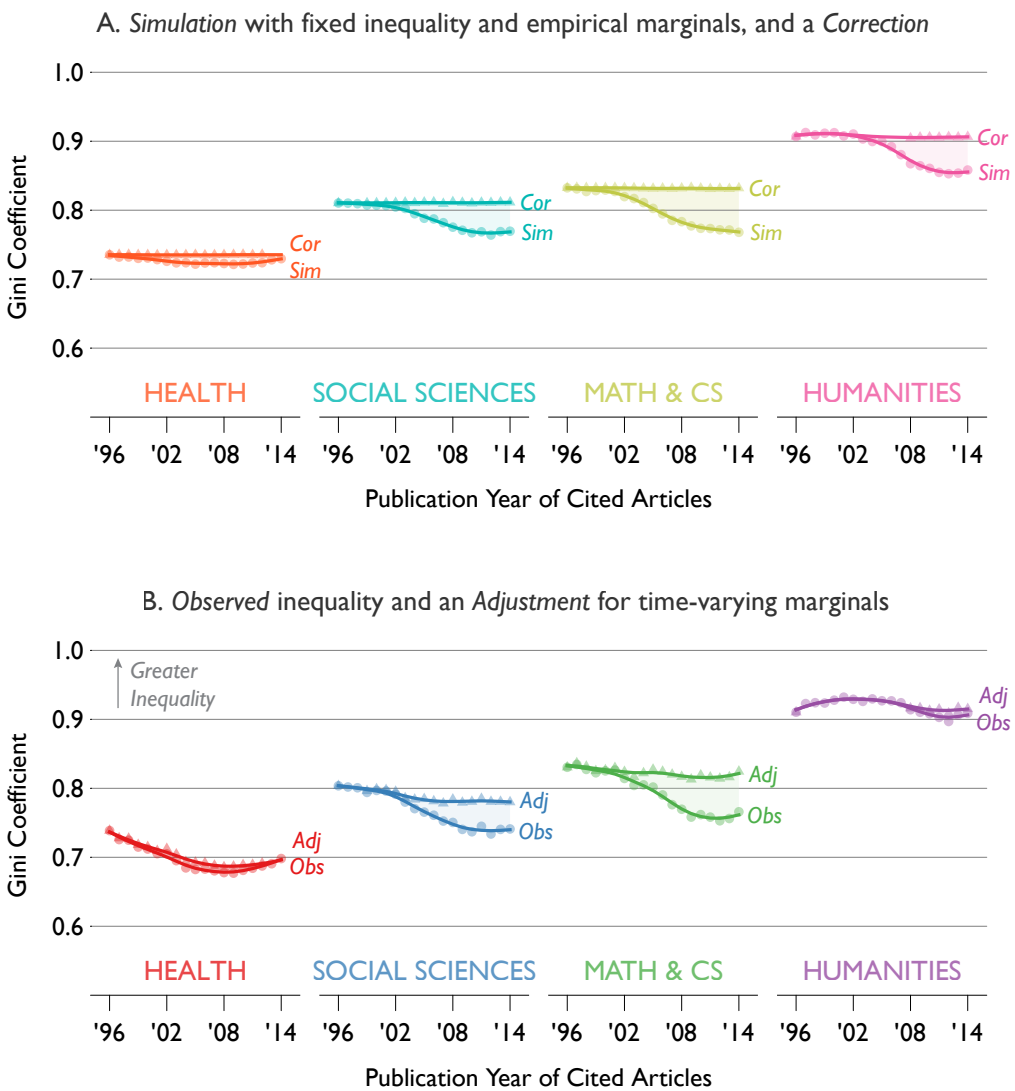

Figure S2. *Gini coefficient for citations within two years of publication with the fixed set of journals continuously published between 1996–2014: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show Gini coefficients of citation distribution from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in the Gini coefficient using a resampling correction. The lines marked *Obs* in the bottom panel shows the Gini coefficients over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the health in 2013–2014, in the social sciences in 1997–1998, and in the humanities in 1997–2007. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

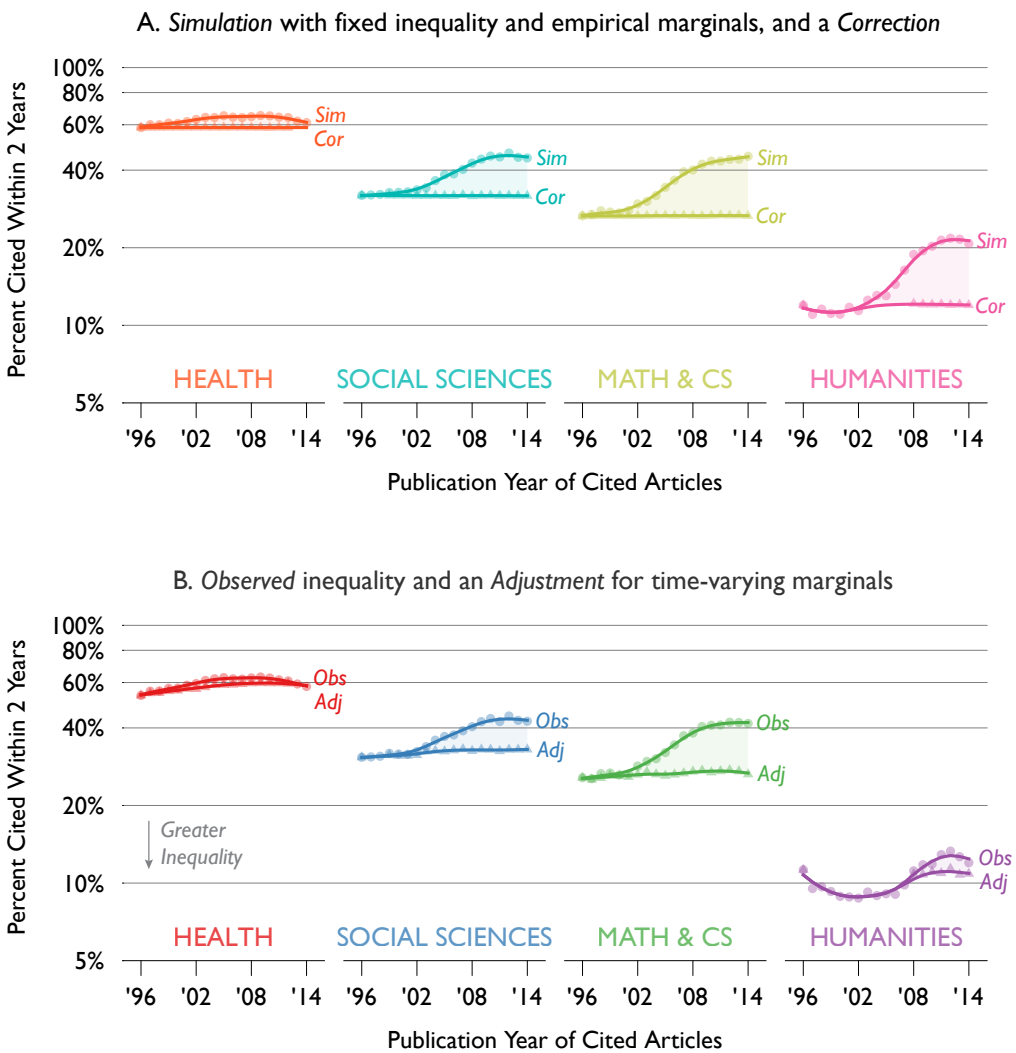

Figure S3. *Percent of papers with any citations within two years of publication with the fixed set of journals continuously published between 1996–2014: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers ever cited from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in percent-ever-cited using a resampling correction. The lines marked *Obs* in the bottom panel shows percent-ever-cited over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the health in 2013–2014, in the social sciences in 1997–1998, and in the humanities in 1997–2007. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

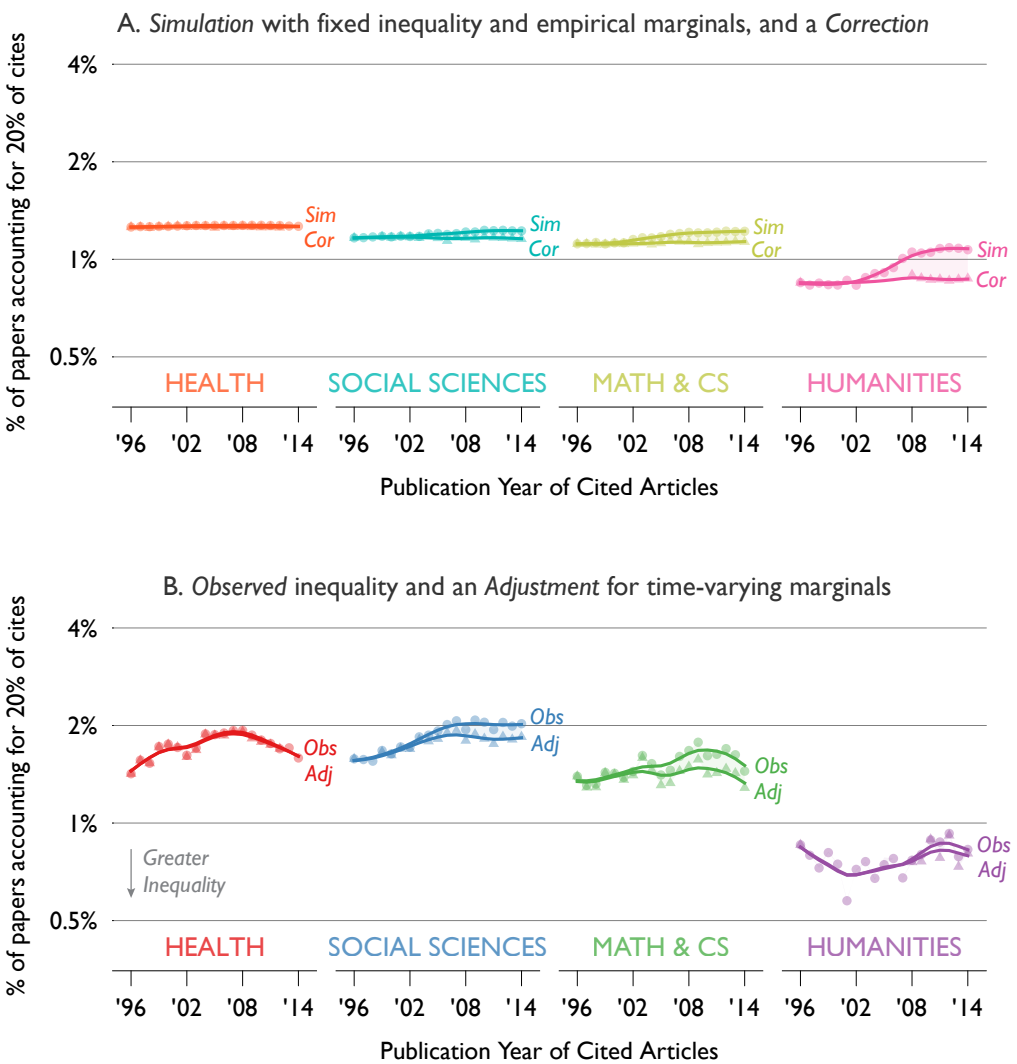

Figure S4. *Percent of papers accounting for 20% of all citations within two years of publication with the fixed set of journals continuously published between 1996–2014: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers accounting for 20% of all citations from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias using a resampling correction. The lines marked *Obs* in the bottom panel shows percent of papers accounting for 20% of citations over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the health in 2013–2014, in the social sciences in 1997–1998, and in the humanities in 1997–2007. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

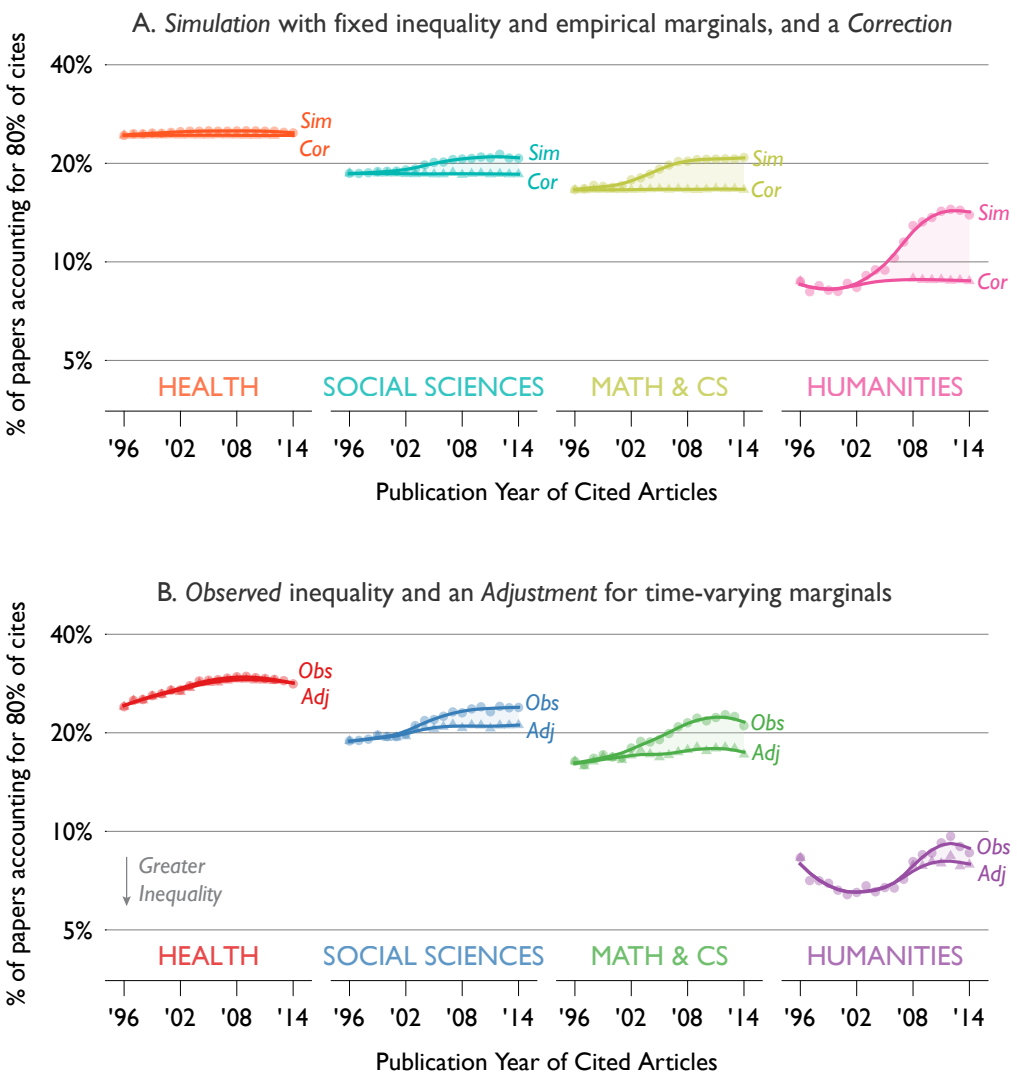

Figure S5. *Percent of papers accounting for 80% of all citations within two years of publication with the fixed set of journals continuously published between 1996–2014: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers accounting for 80% of all citations from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias using a resampling correction. The lines marked *Obs* in the bottom panel shows percent of papers accounting for 80% of citations over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the health in 2013–2014, in the social sciences in 1997–1998, and in the humanities in 1997–2007. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

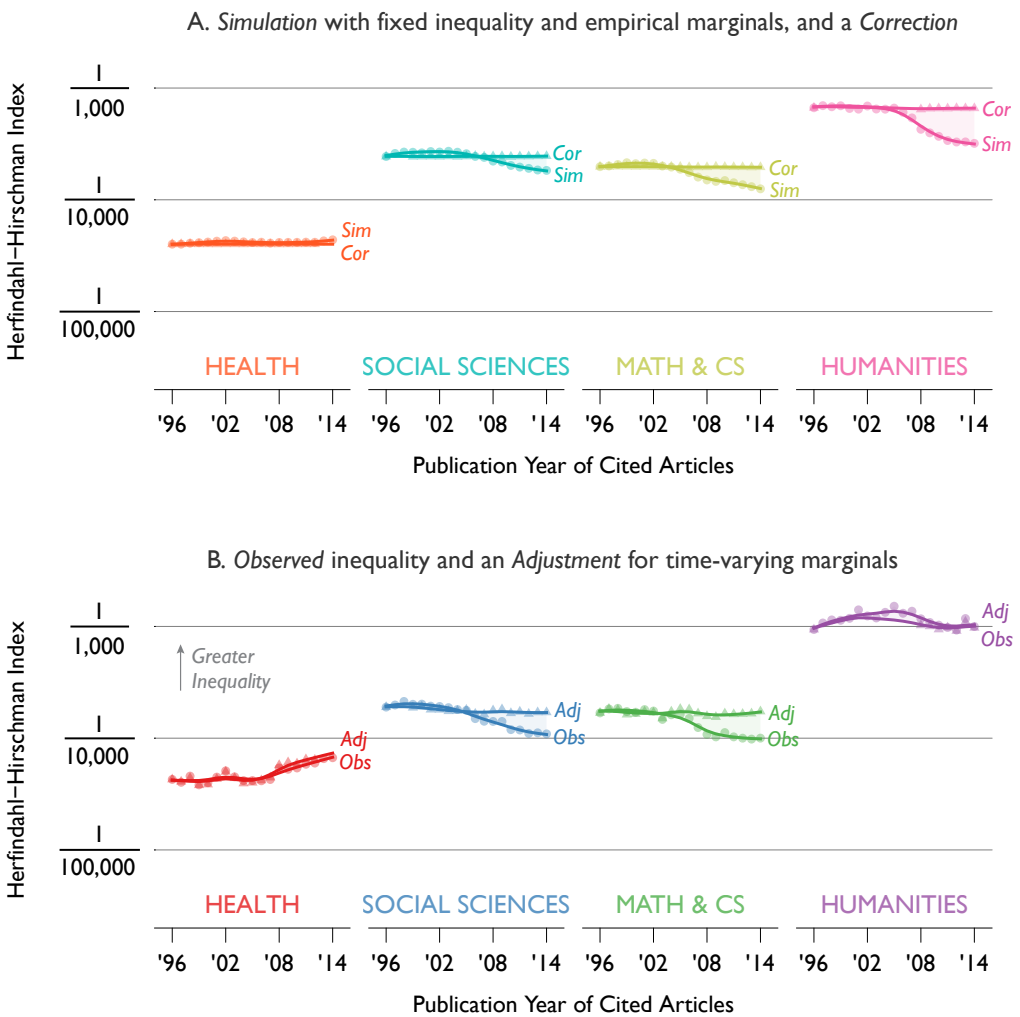

Figure S6. *Herfindahl-Hirschman Index of citations within two years of publication with the fixed set of journals continuously published between 1996–2014: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show the Herfindahl-Hirschman Index (HHI) of citation concentration from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in HHI using a resampling correction. The lines marked *Obs* in the bottom panel shows HHI over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the health in 2013–2014, in the social sciences in 1997–1998, and in the humanities in 1997–2007. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

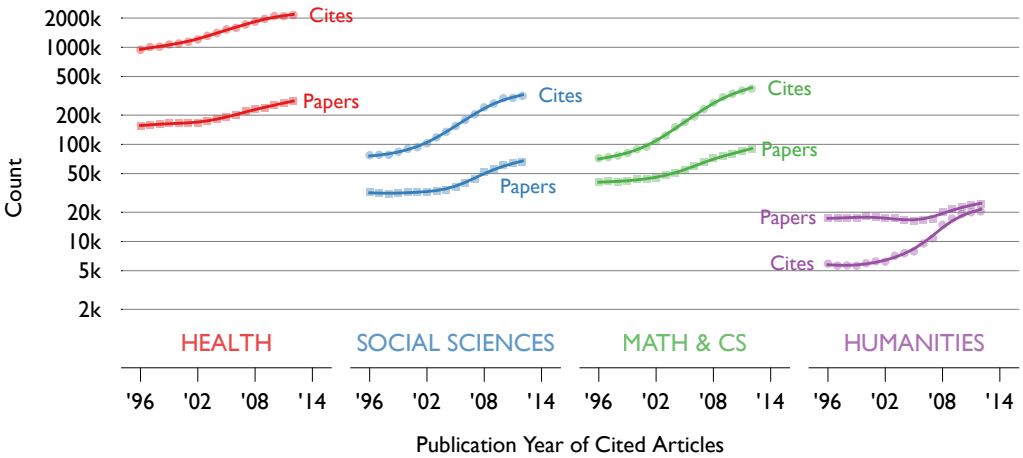

Figure S7. *Number of journal articles published 1996–2012 and citations to those articles within four years of publication.* Compiled from the Web of Science (Clarivate Analytics). Trend lines estimated by robust-and-resistant regression to minimize the influence of outliers. All curves are smoothing splines with span of 0.5.

is also a visual artifact of comparisons across different citation windows. Due to a lack of data past 2016, we cannot report results for four-year windows for papers published after 2012, but these were the publication periods most affected by marginals bias when compared to papers published in 1996. Their omission makes these figures appear less biased because only earlier periods can be compared.)

S 2.3 Analysis with longer time windows: citations over six years

Again, we assess the sensitivity of our results presented in the main text with the two-year citation window, we repeat our analysis using six-year citation windows. Because our analysis includes citations made between 1996 and 2016, the six-year citation window only includes papers published up to 2010 (six years before 2016.) Figure S13 shows the total number of papers published in each of these years, by field, and the total number of citations to those papers in the following six years. Total citations increased in every field except the humanities from 1997–2000, which are the only cases that could not be adjusted for marginals bias.

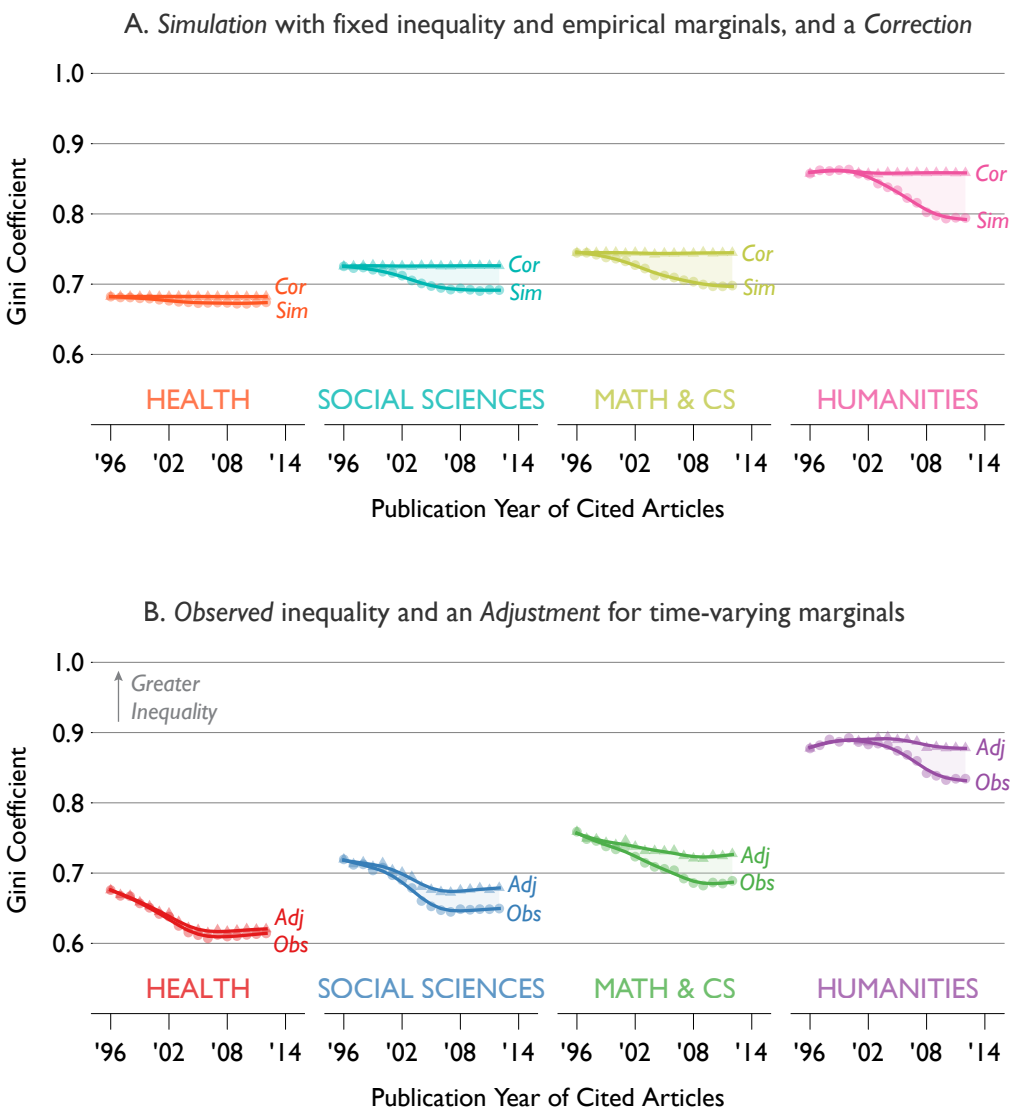

Figure S8. *Gini coefficient for citations within four years of publication, 1996–2012: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show Gini coefficients of citation distribution from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in the Gini coefficient using a resampling correction. The lines marked *Obs* in the bottom panel shows the Gini coefficients over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2000. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

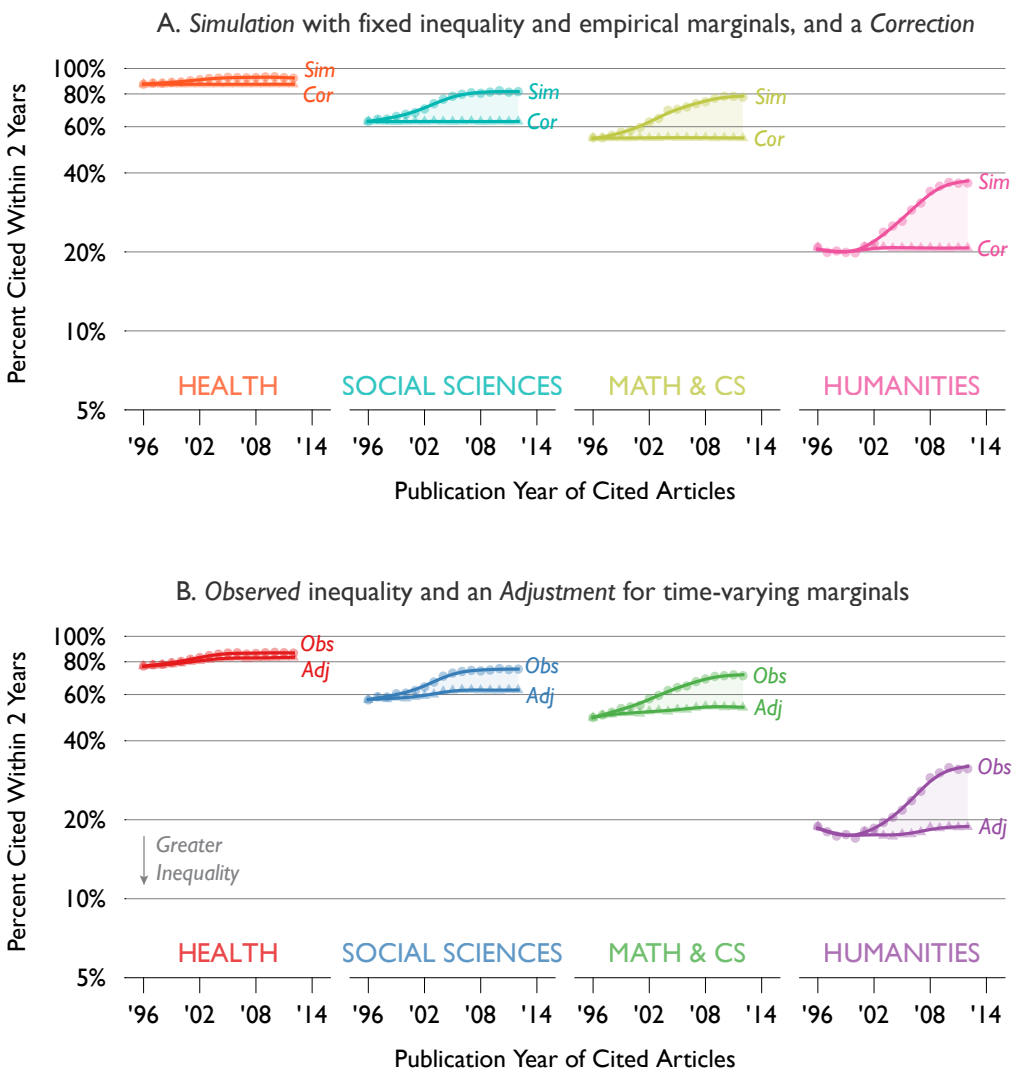

Figure S9. *Percent of papers with any citations within four years of publication, 1996–2012: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers ever cited from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in percent-ever-cited using a resampling correction. The lines marked *Obs* in the bottom panel shows percent-ever-cited over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2000. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

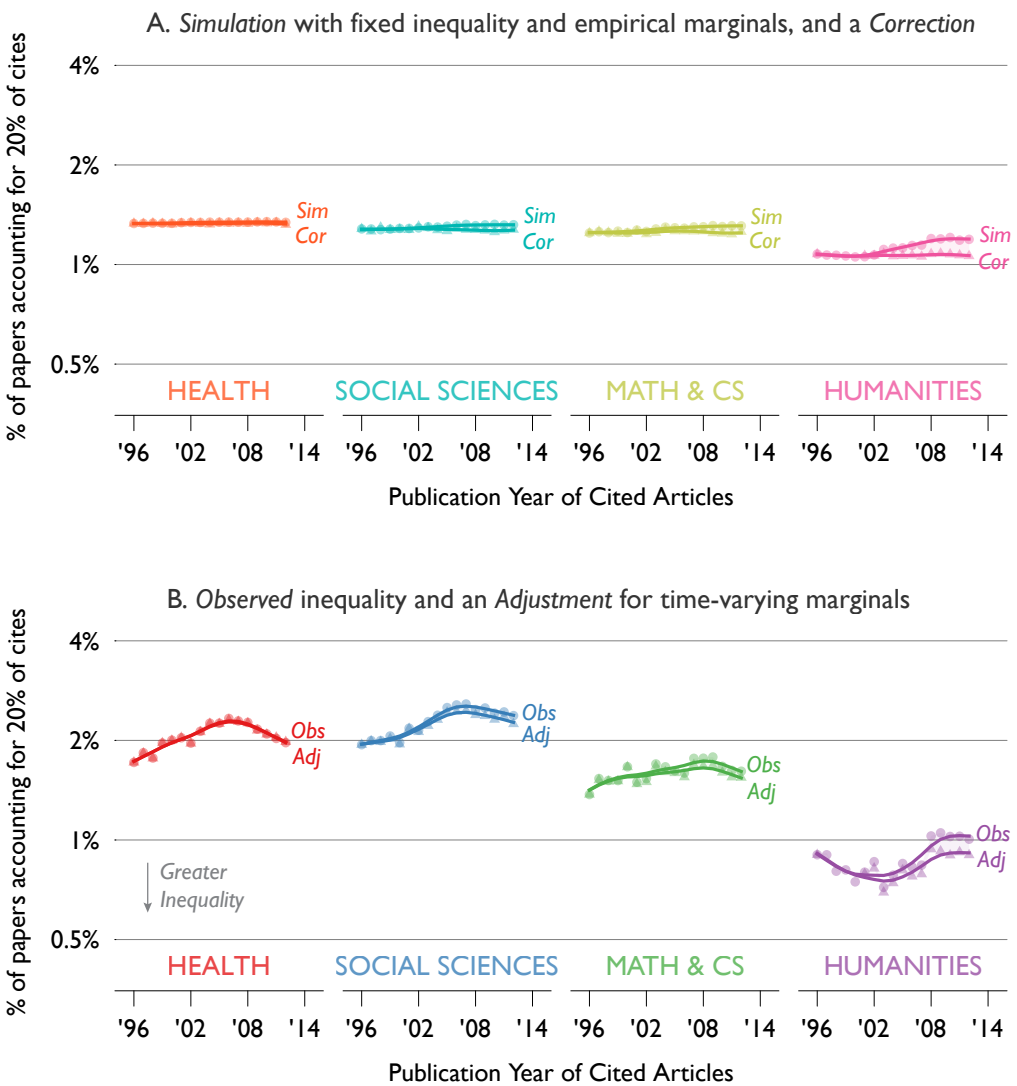

Figure S10. *Percent of papers accounting for 20% of all citations within four years of publication, 1996–2012: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers accounting for 20% of all citations from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias using a resampling correction. The lines marked *Obs* in the bottom panel shows percent of papers accounting for 20% of citations over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2000. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

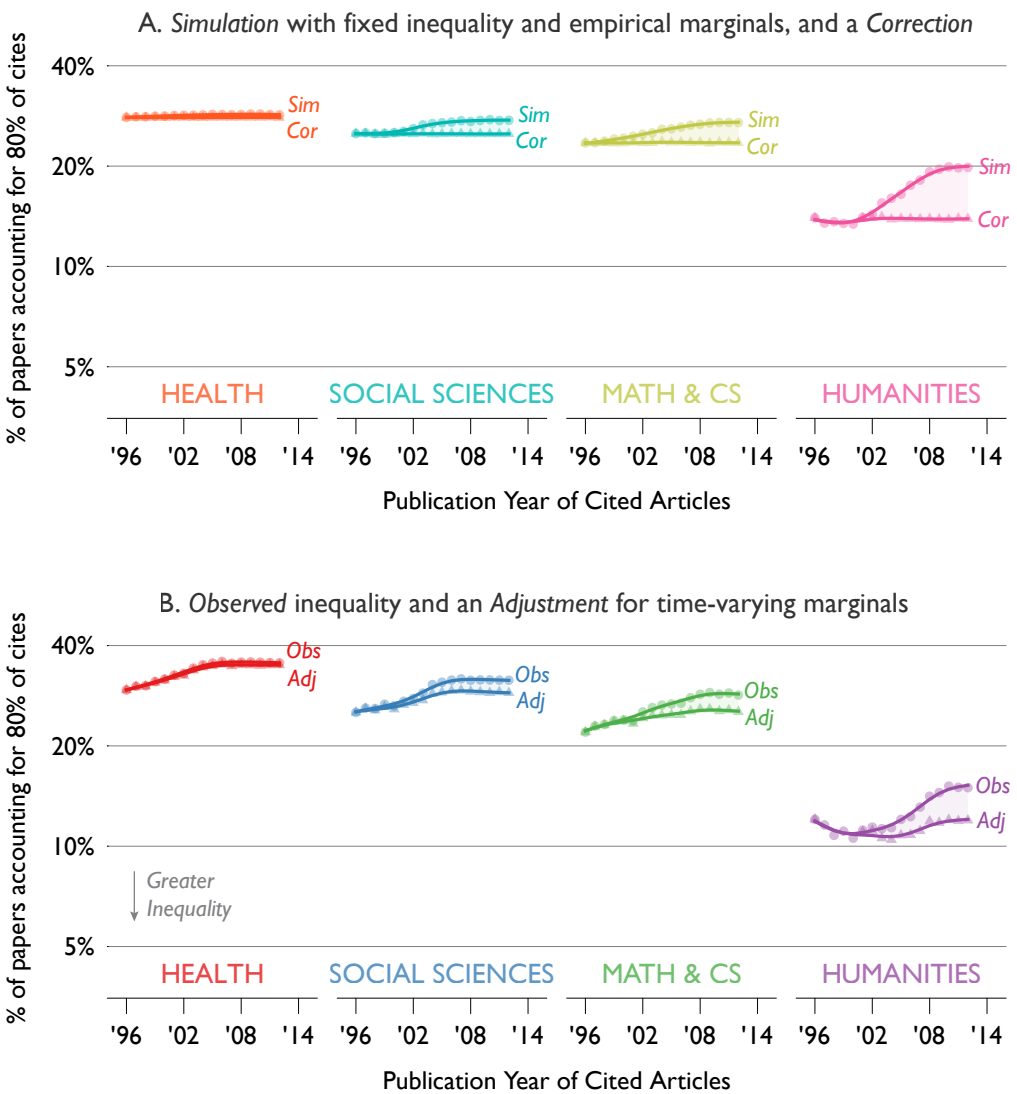

Figure S11. *Percent of papers accounting for 80% of all citations within four years of publication, 1996–2012: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers accounting for 80% of all citations from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias using a resampling correction. The lines marked *Obs* in the bottom panel shows percent of papers accounting for 80% of citations over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2000. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

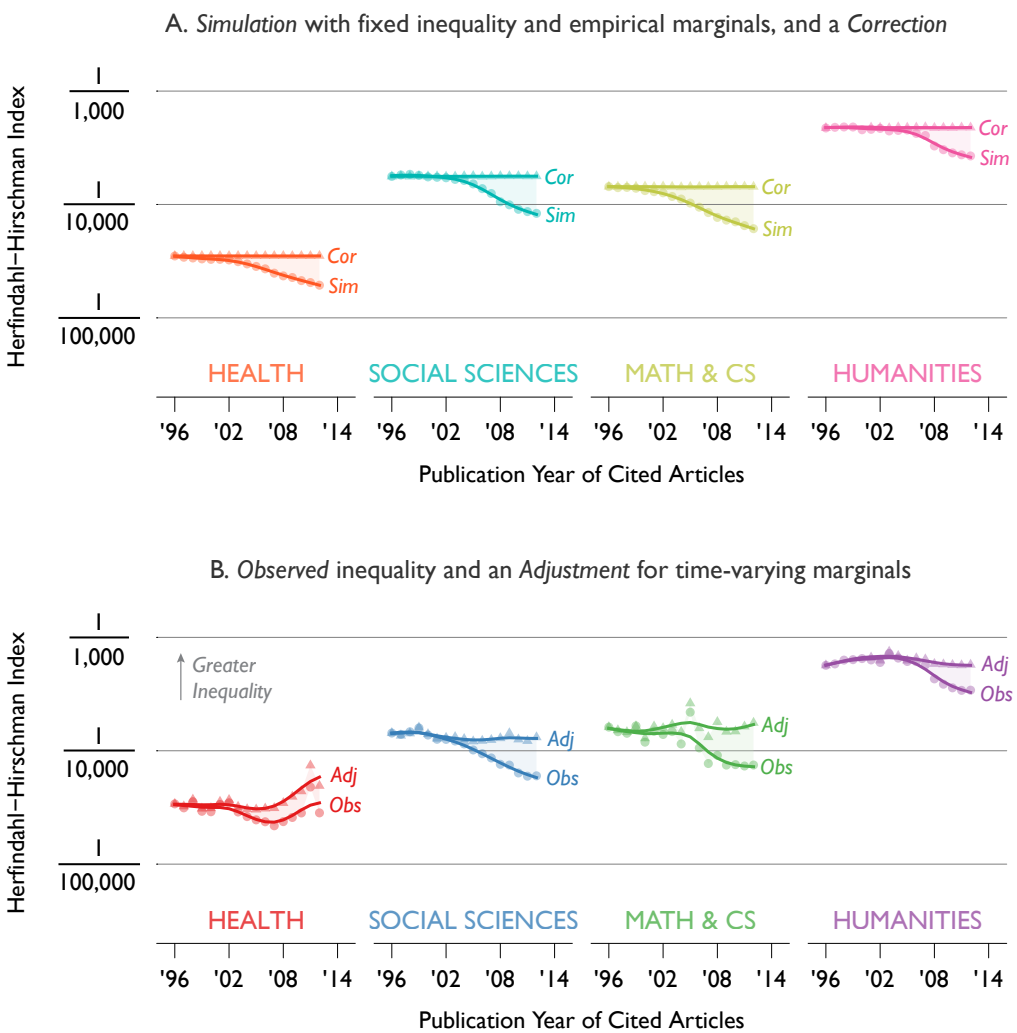

Figure S12. *Herfindahl-Hirschman Index of citations within four years of publication, 1996–2012: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show the Herfindahl-Hirschman Index (HHI) of citation concentration from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in HHI using a resampling correction. The lines marked *Obs* in the bottom panel shows HHI over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2000. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

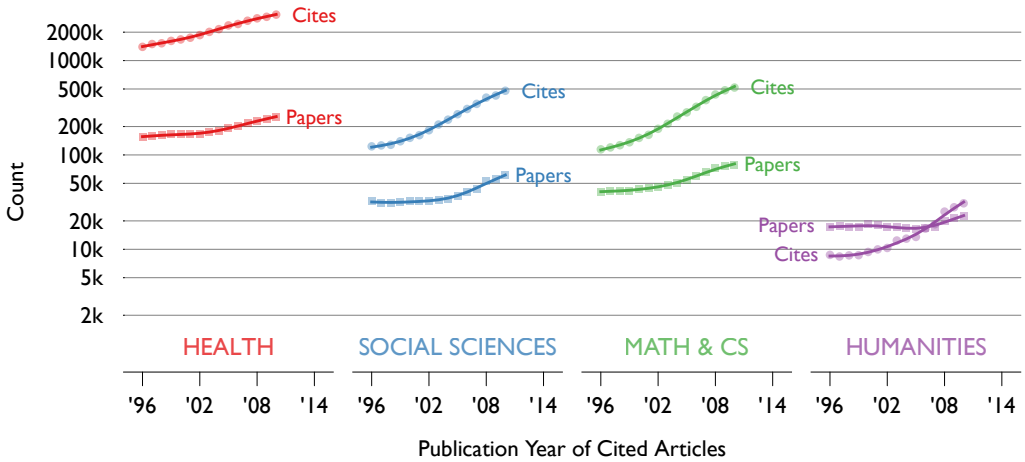

Figure S13. *Number of journal articles published 1996–2010 and citations to those articles within six years of publication.* Compiled from the Web of Science (Clarivate Analytics). Trend lines estimated by robust-and-resistant regression to minimize the influence of outliers. All curves are smoothing splines with span of 0.5.

The changes in citation concentration over time observed follow patterns similar to those produced using the shorter window in analyses presented in the main text (Figures S14–S18). The degree of marginals bias grows smaller as the citation window grows longer – and fewer published papers thus remain close to the zero-lower bound for citations – but does not disappear, suggesting adjustment is still necessary for longer windows.

Part of the apparent reduction of marginals bias is also a visual artifact of comparisons across different citation windows. Due to a lack of data past 2016, we cannot report results for six-year windows for papers published after 2010, but these were the publication periods most affected by marginals bias when compared to papers published in 1996. Their omission makes these figures appear less biased because only earlier periods can be compared.

S 2.4 Analysis including a single outlier paper in mathematics and computer sciences

In our results in the main text, we omit a single unusually highly-cited paper in mathematics and computer sciences. This section shows what happens to our main 2-year

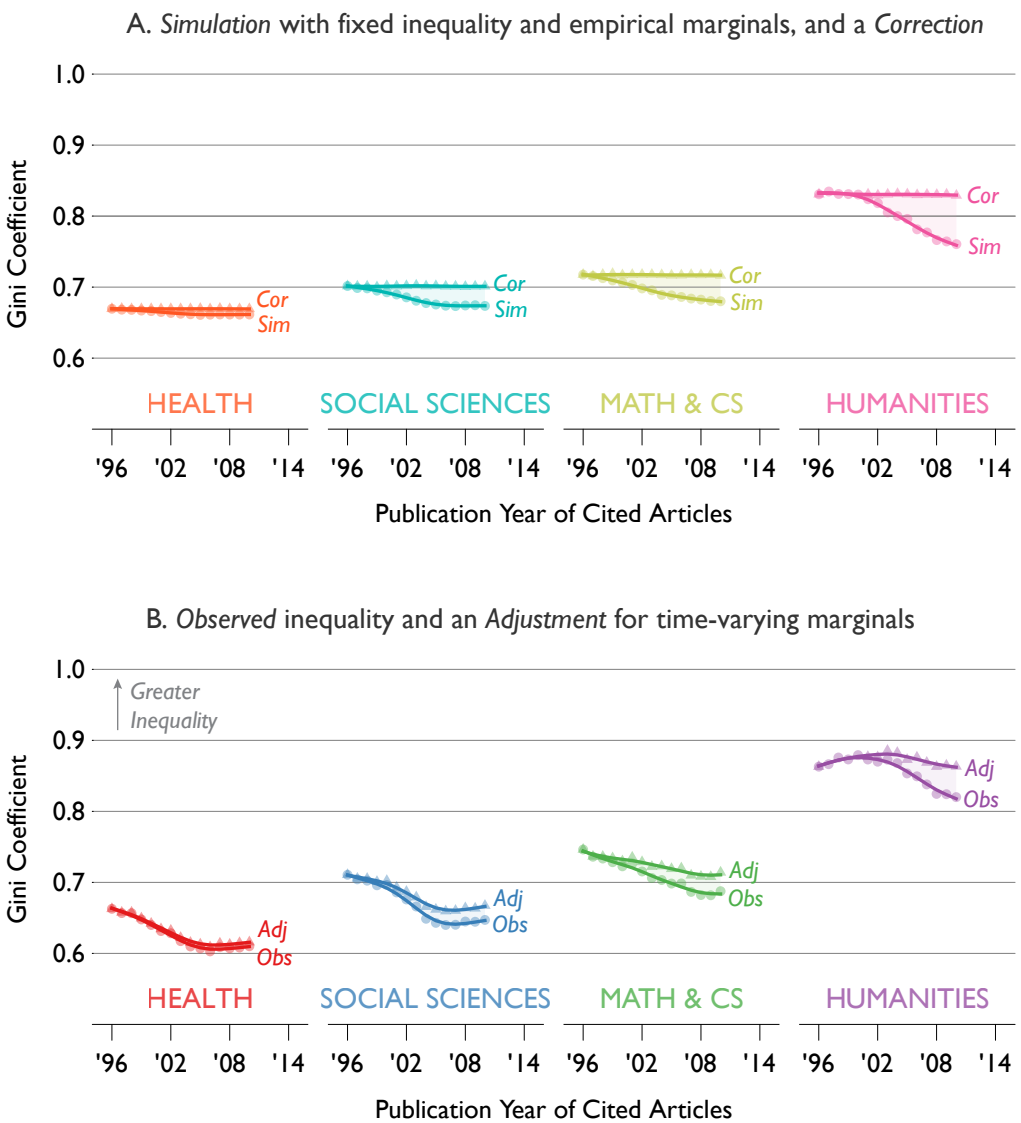

Figure S14. *Gini coefficient for citations within six years of publication, 1996–2010: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show Gini coefficients of citation distribution from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in the Gini coefficient using a resampling correction. The lines marked *Obs* in the bottom panel shows the Gini coefficients over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2000. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

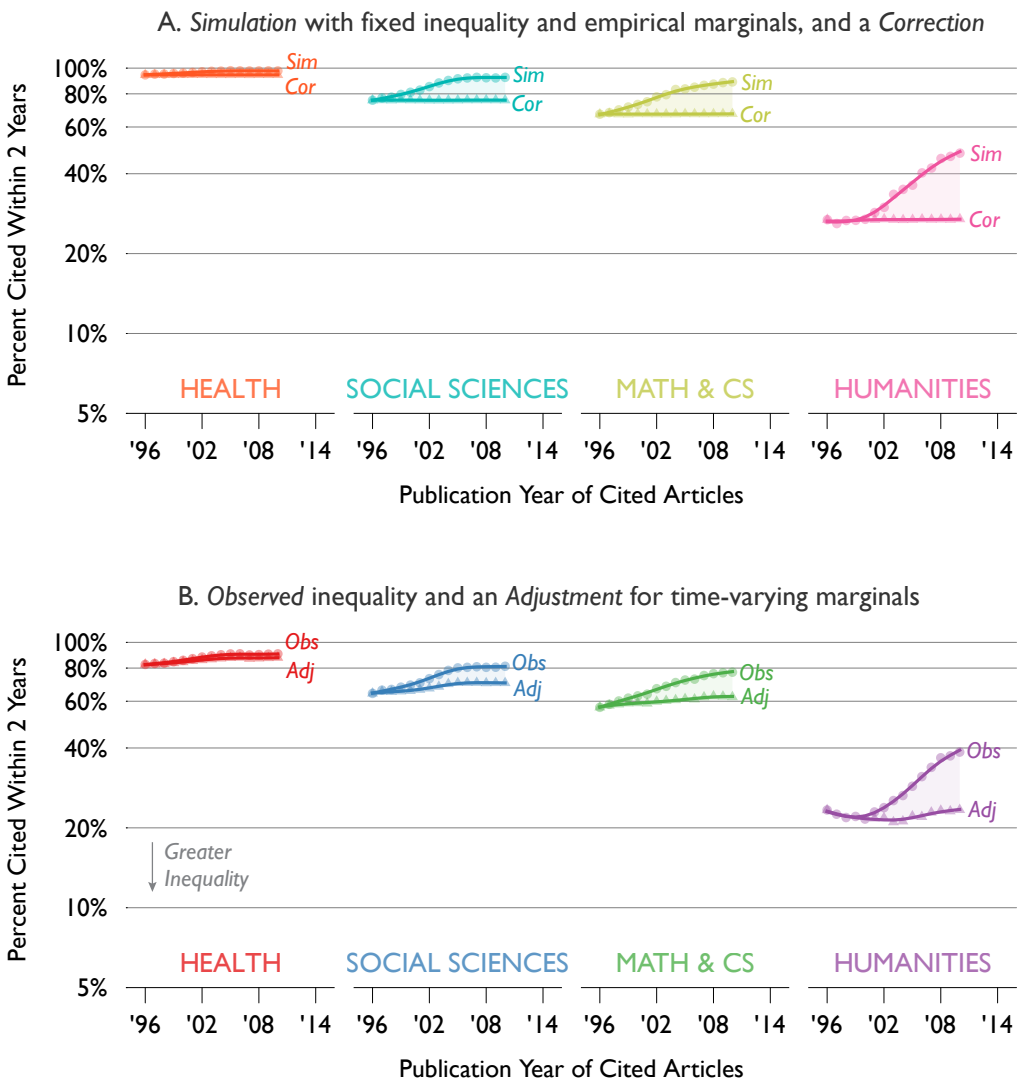

Figure S15. *Percent of papers with any citations within six years of publication, 1996–2010: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers ever cited from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in percent-ever-cited using a resampling correction. The lines marked *Obs* in the bottom panel shows percent-ever-cited over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2000. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

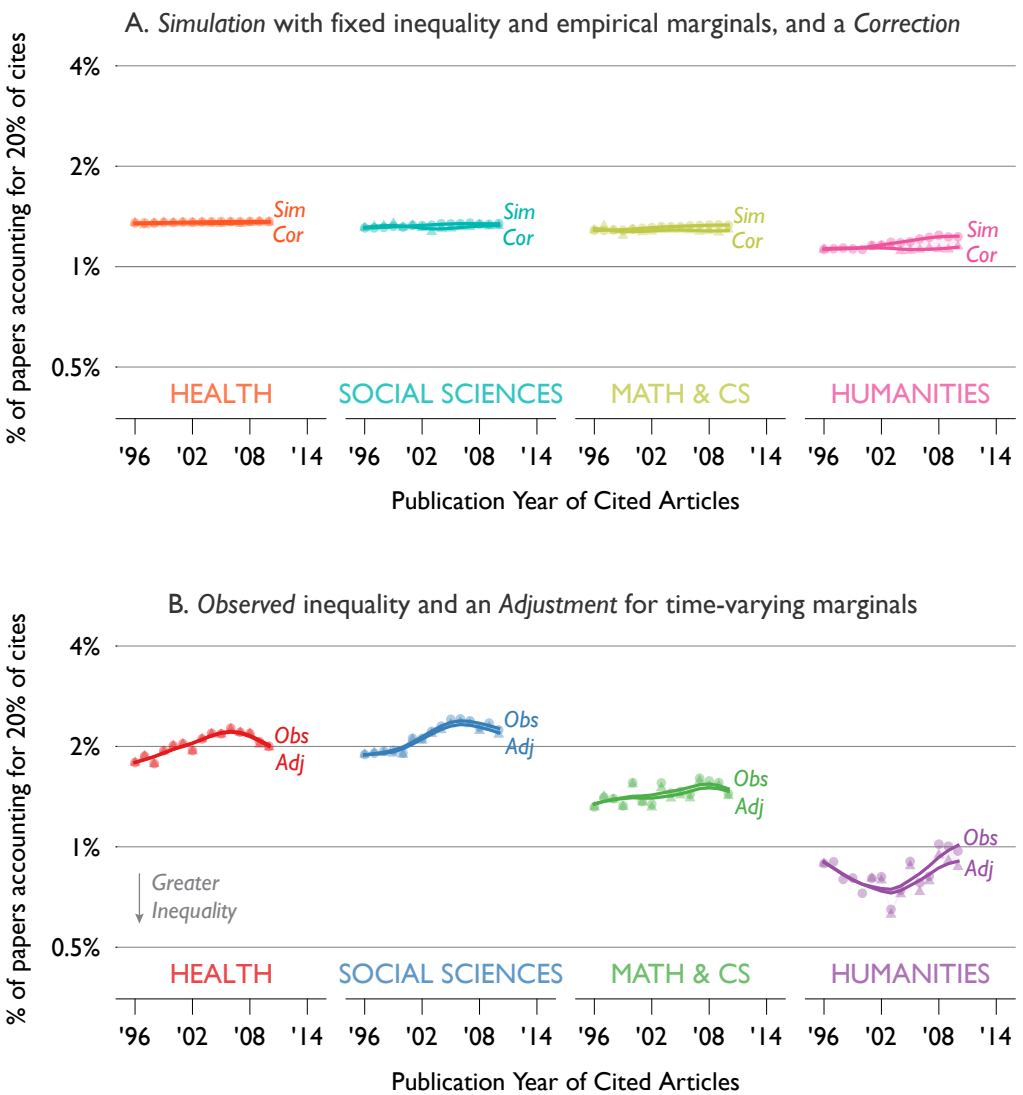

Figure S16. *Percent of papers accounting for 20% of all citations within six years of publication, 1996–2010: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers accounting for 20% of all citations from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias using a resampling correction. The lines marked *Obs* in the bottom panel shows percent of papers accounting for 20% of citations over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2000. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

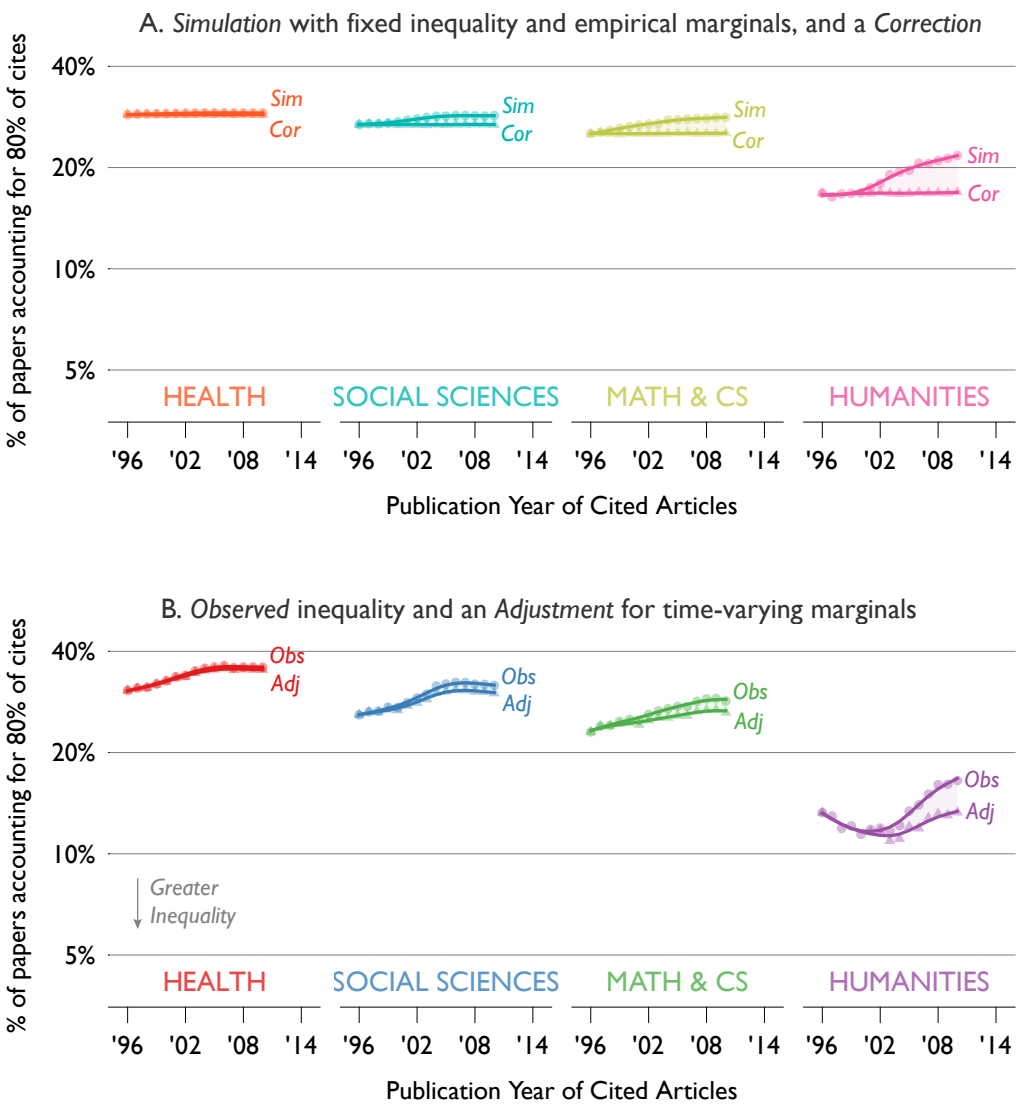

Figure S17. *Percent of papers accounting for 80% of all citations within six years of publication, 1996–2010: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers accounting for 80% of all citations from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias using a resampling correction. The lines marked *Obs* in the bottom panel shows percent of papers accounting for 80% of citations over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2000. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

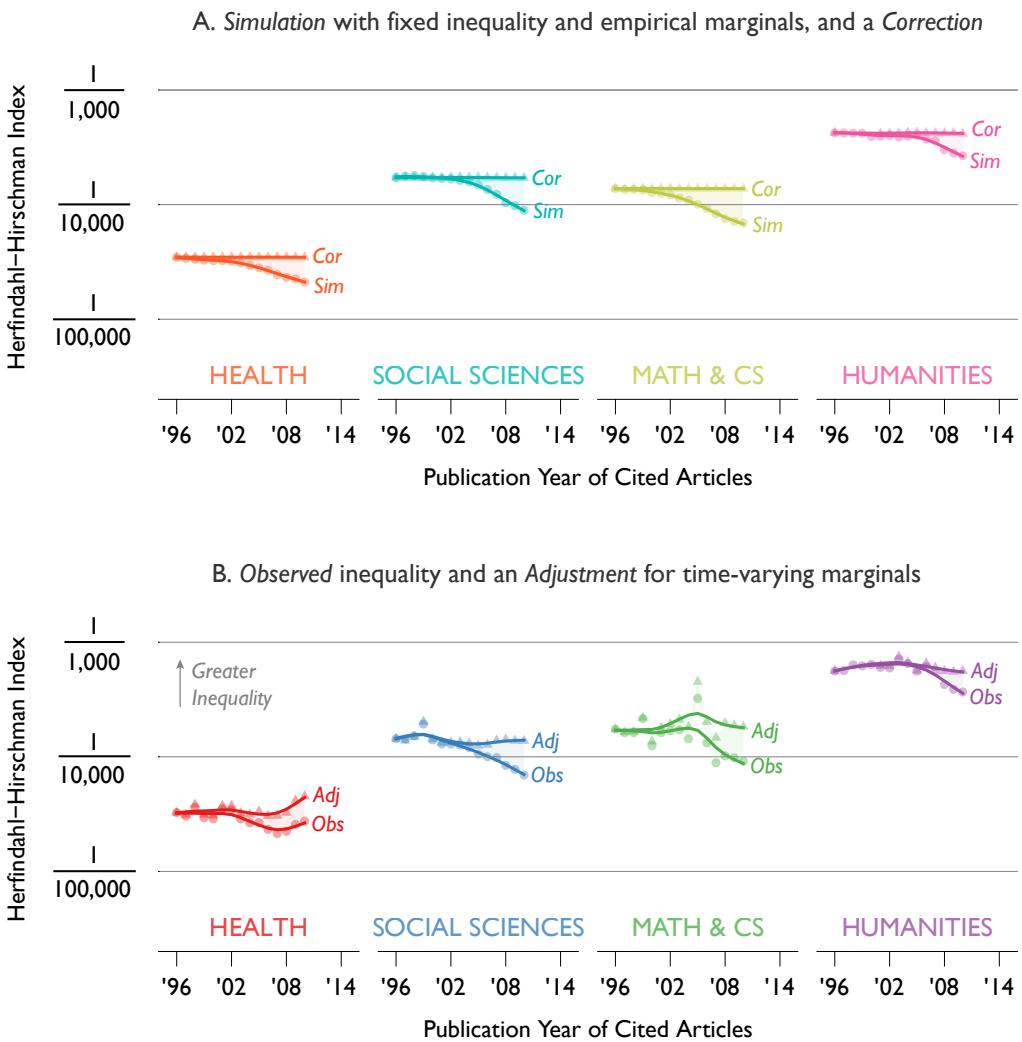

Figure S18. *Herfindahl-Hirschman Index of citations within six years of publication, 1996–2010: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show the Herfindahl-Hirschman Index (HHI) of citation concentration from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in HHI using a resampling correction. The lines marked *Obs* in the bottom panel shows HHI over fields and time using the empirical data from World of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2000. All curves are smoothing splines with span of 0.5. One exceptionally highly cited paper in Math & CS is omitted.

window results when we include this paper. With the exception of the Herfindahl-Hirschman Index, including the outlier makes little or no discernable difference (compare the Math & CS plots in the bottom row of main text Figures 3–6 with the corresponding plots in Figures S19-S22 below). However, because Herfindahl-Hirschman indexes are particularly sensitive to extreme cases of concentration, including this single paper produces a strong outlier in the HHI results (compare Figure 7 in the main text to Figure S23 below). The degree to which HHI is influenced by this single outlier is unaffected by our resampling correction.

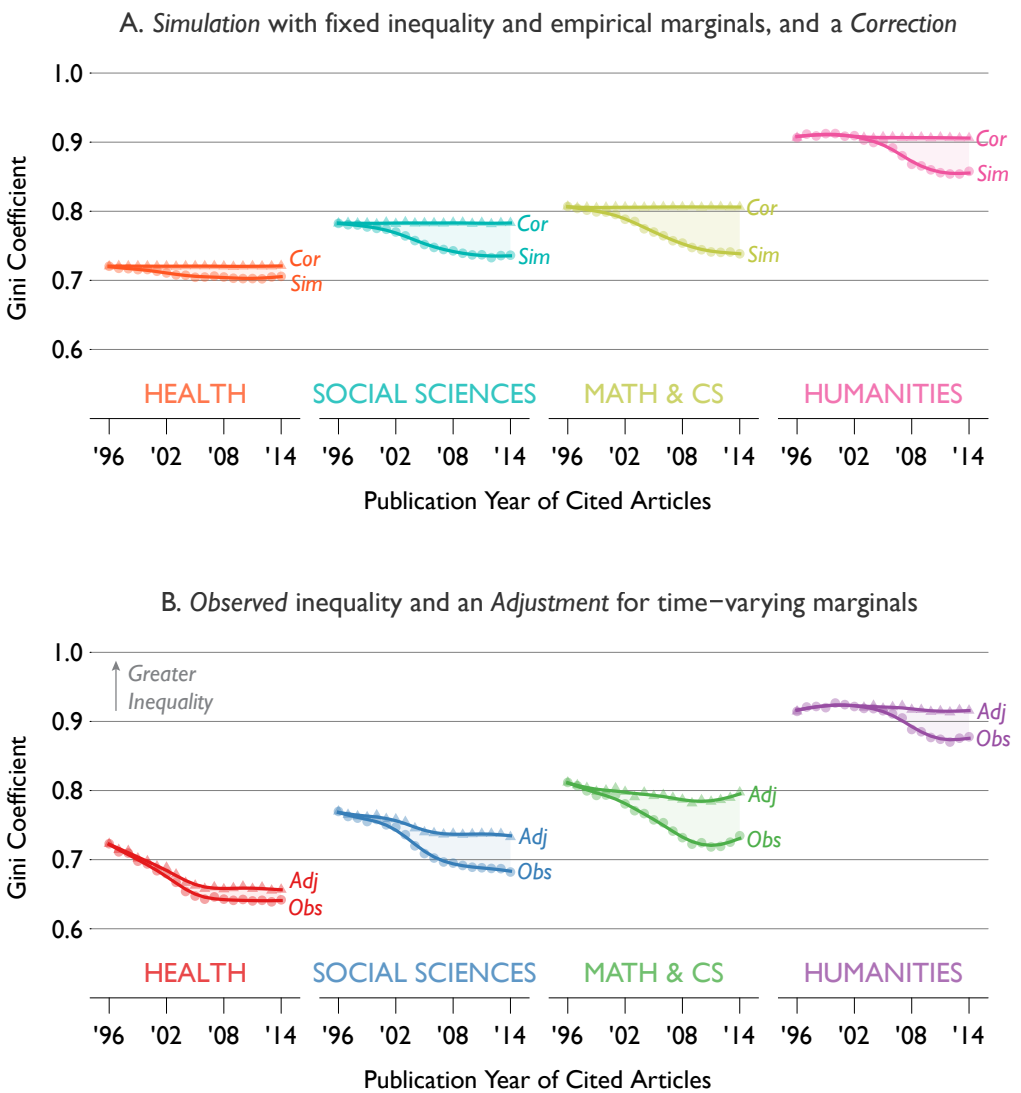

Figure S19. *Gini coefficient for citations within two years of publication, 1996–2014, including an outlier: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show Gini coefficients of citation distributions from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in the Gini coefficient using a resampling correction. The lines marked *Obs* in the bottom panel shows Gini coefficients over fields and time using the empirical data from Web of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2002. All curves are smoothing splines with span of 0.5. No cases are omitted.

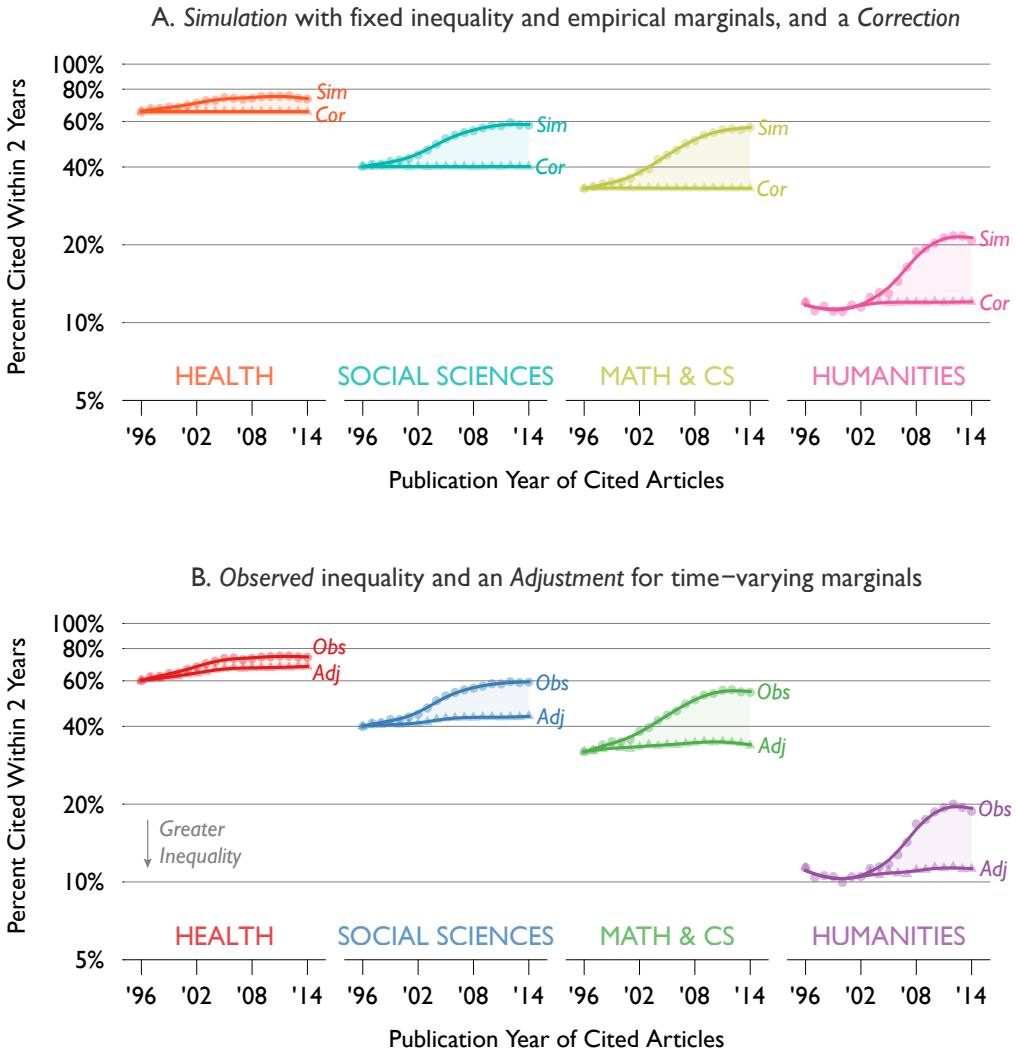

Figure S20. *Percent of papers with any citations two years after publication, 1996–2014, including an outlier: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers ever cited from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in percent-ever-cited using a resampling correction. The lines marked *Obs* in the bottom panel shows percent-ever-cited over fields and time using the empirical data from Web of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2002. All curves are smoothing splines with span of 0.5. No cases are omitted.

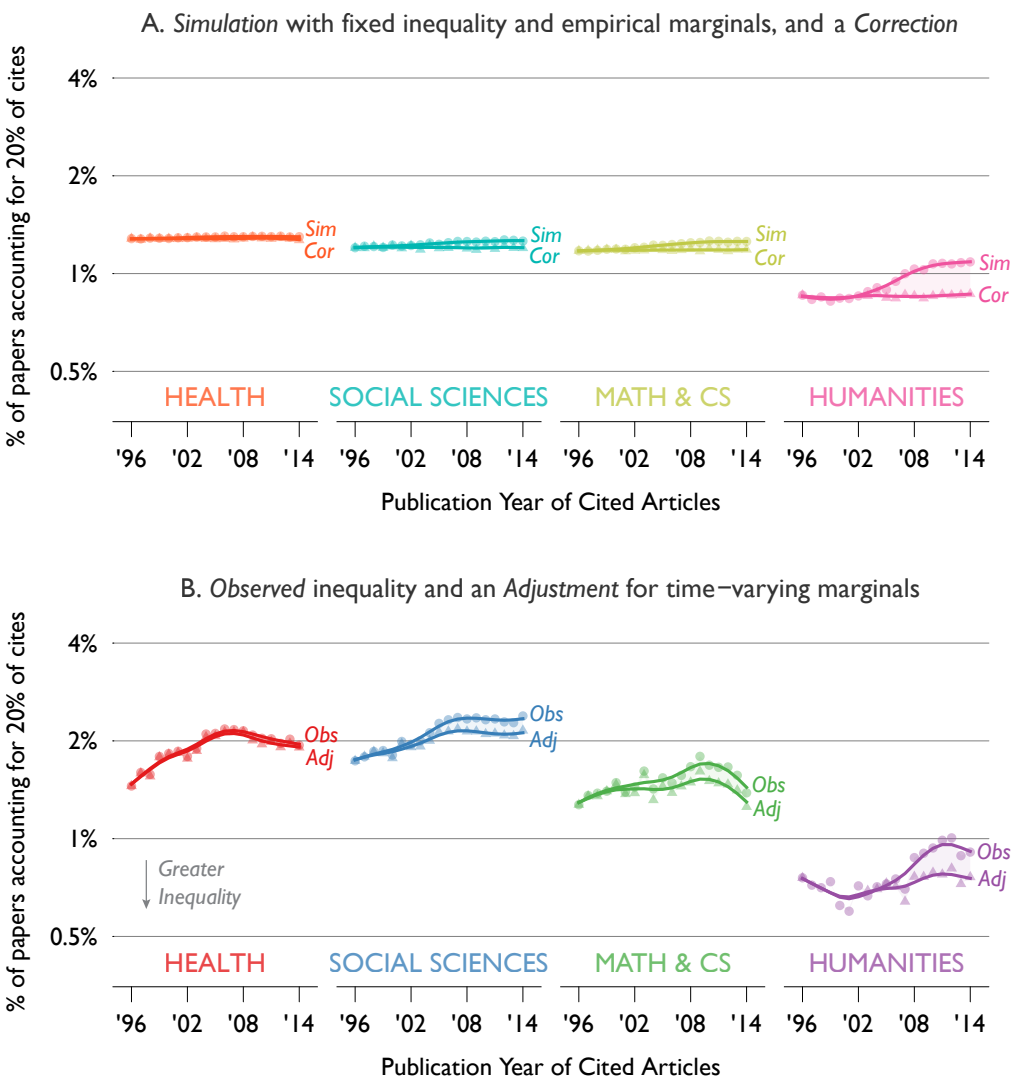

Figure S21. *Percent of papers accounting for 20% of all citations within two years of publication, 1996–2014, including an outlier: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers accounting for 20% of all citations from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias using a resampling correction. The lines marked *Obs* in the bottom panel shows the percent of papers accounting for 20% of citations over fields and time using the empirical data from Web of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2002. All curves are smoothing splines with span of 0.5. No cases are omitted.

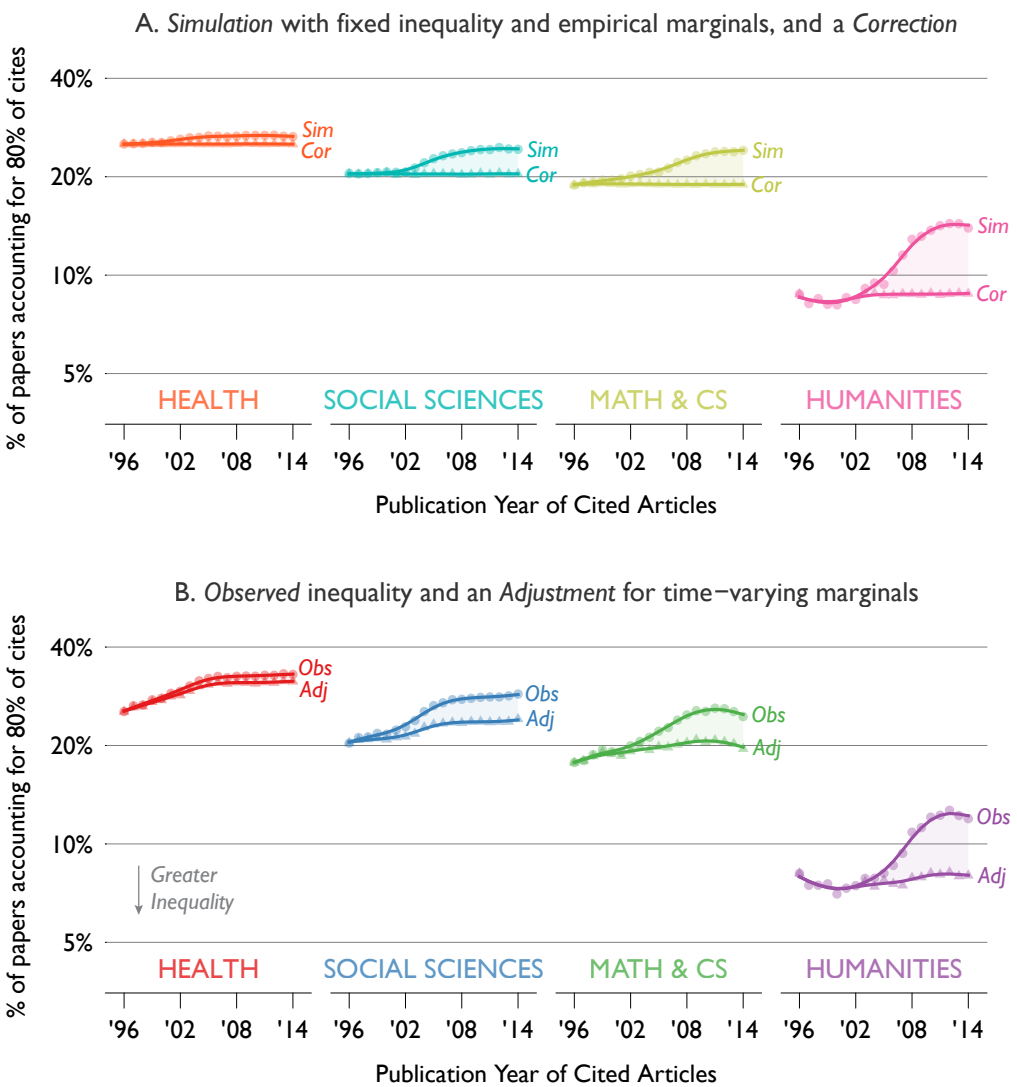

Figure S22. *Percent of papers accounting for 80% of all citations within two years of publication, 1996–2014, including an outlier: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show percent of papers accounting for 80% of all citations from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias using a resampling correction. The lines marked *Obs* in the bottom panel shows the percent of papers accounting for 80% of citations over fields and time using the empirical data from Web of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2002. All curves are smoothing splines with span of 0.5. No cases are omitted.

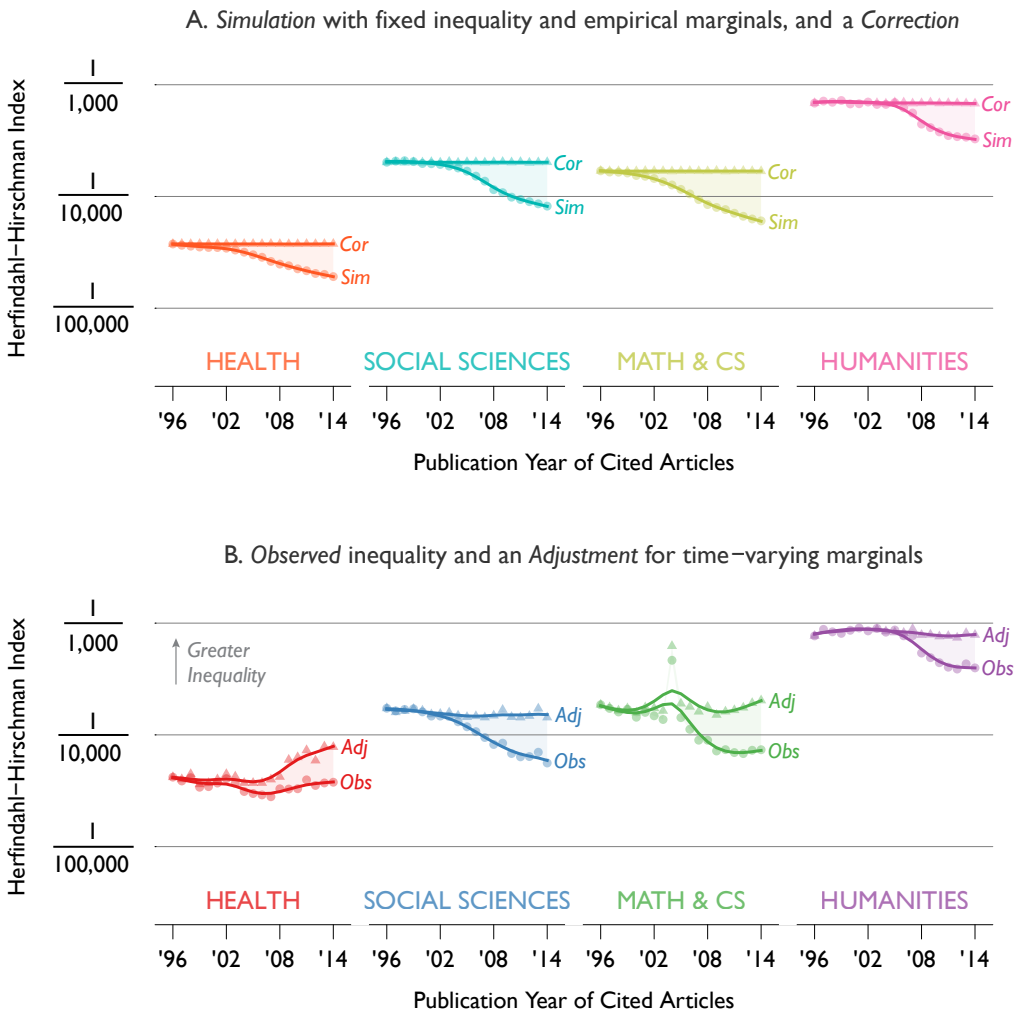

Figure S23. *Herfindahl-Hirschman Index of citations within two years of publication, 1996–2014, including an outlier: Monte Carlo simulation and empirical results.* The lines in the top panel marked *Sim* show the Herfindahl-Hirschman Index (HHI) of citation concentration from Monte Carlo results for hypothetical papers and citations designed to have a fixed pattern of inequality across years and fields, but total papers and citations matching the empirical marginals of those fields by year. The lines marked *Cor* remove the marginals bias in HHI using a resampling correction. The lines marked *Obs* in the bottom panel shows HHI over fields and time using the empirical data from Web of Science; these results are subject to marginals bias from differences in total papers and citations by field and year. Lines marked *Adj* adjust for marginals bias in the empirical data by resampling to the marginals in 1996 by field. Corrections and adjustments omitted for the humanities in 1997–2002. All curves are smoothing splines with span of 0.5. No cases are omitted.