Dataset Artefacts are the Hidden Drivers of the Declining Disruptiveness in Science

Vincent Holst[1*], Andres Algaba[1], Floriano Tori[1], Sylvia Wenmackers[2], Vincent Ginis[1,3*]

[1]Data Analytics Laboratory, Vrije Universiteit Brussel, Pleinlaan 2, Brussels, 1050, Belgium.
[2]Centre for Logic and Philosophy of Science (CLPS), KU Leuven, K. Mercierplein, 2 – box 3200, Leuven, 3000, Belgium.
[3]School of Engineering and Applied Sciences, Harvard University, 9 Oxford Street, Boston, MA 02134, Massachusetts, USA.

*Corresponding author(s). E-mail(s): vincent.thorge.holst@vub.be; vincent.ginis@vub.be;
Contributing authors: andres.algaba@vub.be; floriano.tori@vub.be; sylvia.wenmackers@kuleuven.be;

Park et al. [1] reported a decline in the disruptiveness of scientific and technological knowledge over time. Their main finding is based on the computation of CD indices, a measure of disruption in citation networks [2], across almost 45 million papers and 3.9 million patents. Due to a factual plotting mistake, database entries with zero references were omitted in the CD index distributions, hiding a large number of outliers with a maximum CD index of one, while keeping them in the analysis [1]. Our reanalysis shows that the reported decline in disruptiveness can be attributed to a relative decline of these database entries with zero references. Notably, this was not caught by the robustness checks included in the manuscript. The regression adjustment fails to control for the hidden outliers as they correspond to a discontinuity in the CD index. Proper evaluation of the Monte-Carlo simulations reveals that, because of the preservation of the hidden outliers, even random citation behaviour replicates the observed decline in disruptiveness. Finally, while these papers and patents with supposedly zero references are the hidden drivers of the reported decline, their source documents predominantly do make references, exposing them as pure dataset artefacts.

Fig. 1**a,d** reproduces the CD_5 index distributions for papers and patents as presented in Park et al. [1] (Extended Data Fig. 1**a,c** in [1]). A bug in the *seaborn 0.11.2* plotting software [3], used by Park et al. [1], silently drops the largest data points in the histograms. Therefore, these histograms do not show the papers and patents with $CD_5 = 1$. Using correct plot settings, Fig. 1**b,e** reveals the additional $972,161$ papers and $142,362$ patents, with $CD_5 = 1$. However, these hidden outliers were included in the main analysis in [1]: the evaluation of the disruption versus time. Fig. 1**c,f** shows that the decline in the disruptiveness of scientific (resp. technological) knowledge over time is negated (resp. substantially reduced) when these outliers are excluded.

The origin of these data points, and our reason for calling them outliers, can be found in their metadata. For patents, these are publicly available in the *PatentsView* data source. As an open source alternative for *Web of Science*, we use *SciSciNet* [4], an equivalent citation network with $39,888,199$ papers, for which we could replicate the above observation that $CD_5 = 1$ papers are responsible for the temporal decline in disruption (Extended Data Fig. A1). Extended Data Fig. A2**a,d** shows the number of references made in each of those papers or patents according to the data source. For *SciSciNet* and *PatentsView*, we find that 97% and 78% of the $CD_5 = 1$ papers and patents make zero references, respectively. Extended Data Fig. A2**b,e** shows that within the $CD_5 = 1$ category, the proportion of patents and papers that makes zero references is stable over time. Importantly, Extended Data Fig. A2**c,f** reveals that the relative frequency of patents and papers with zero references and $CD_5 = 1$ decreases over time, mirroring the decline in disruptiveness reported in [1]. A second proof confirming the above mechanism is found by removing papers (Extended Data Fig. A1**c**) and patents (Fig. 1**f**) with $CD_5 = 1$ and zero references, which has a similar effect on the reported decline compared to removing all hidden $CD_5 = 1$ outliers, making the decline of disruptive science (technology) disappear (decrease).

One could argue that the consequences of the plotting mistake should have been caught by the robustness checks aimed at controlling for changing citation patterns over time [1]. We show below why both the regression adjustment and the Monte-Carlo simulations failed to do so.

First, Park et al. [1] proposed a linear regression to estimate an adjusted CD_5 (models 4 and 8 in Supplementary Table 1 and Extended Data Fig. 8**b,e** in [1]). The regression aims to predict the marginal effect of time by controlling for the number of references on the paper/patent level, together with fixed effects and additional control variables at the (sub-)field and year level. Notably, in the case of zero references to prior work, the CD index either equals the maximum value of one (when there is at least one citation) or remains undefined (when there are no citations). Fig. 2**a,b** shows that the linear regression model [1] fails to control for this discontinuous effect of zero references. Fig. 2**c,d** confirms that the regression errors (RMSE) peak at zero references. To control for the discontinuity, we extend the regression model from Park et al. [1] by including a dummy variable for zero references (Supplementary Table S1). This results in an improved model fit, quantified by an adjusted explained variance (R^2) increasing from 0.10 to 0.52 for patents and from 0.15 to 0.95 for papers. This improvement is not matched by the inclusion of a dummy variable for any other number of references (insets in Fig. 2**c,d**). Supplementary Fig. S1 shows that by explicitly controlling for

the discontinuous effect of zero references, the decline in the disruptiveness of scientific (technological) knowledge is negated (reduced).

Second, Park et al. [1] conducted Monte Carlo simulations to control for the part of the observed decline caused by the general structure of the citation networks. They used a random rewiring algorithm [5], which preserves the publication years of the involved papers and patents and their number of forward citations and references, but randomly rearranges the citations between them. Based on an average z score (Extended Data Fig. 8**c,f** in [1]; see Supplementary Equation S5 for more details), Park et al. [1] state that "the observed CD_5 values are lower than those from the simulated networks [...] and the gap is widening." However, the temporal evolution of the CD index for the randomly rewired networks in Fig. 2**e,f** instead shows a decline in disruption that almost perfectly mirrors the observed decline from the unaltered databases (Supplementary Fig. S2, S3, S5). Moreover, the gap between the observed CD_5 values and those from the simulated networks narrows over time. The fact that the decline in disruption is present even in the randomly rewired networks can be explained by the degree-preserving nature of the rewiring algorithm, which induces a one-to-one correspondence between zero reference papers/patents in the original and rewired network, thus driving the decline in both networks. Consequently, random citation behaviour provides yet another proof that the relative decrease of zero reference patents and papers with $CD_5 = 1$ per year (Extended Data Fig. A2**c,f**) drives the reported decline in disruption.

Finally, we elucidate the source of this large quantity of papers and patents without references by inspecting 100 randomly extracted patents and papers with zero references and $CD_5 = 1$. We find that 98% of the patent sample and 93% of the paper sample do make references in their original PDF, indicating that most of the $CD_5 = 1$ patents and papers with zero references should be treated as artefacts of the respective data sources rather than meaningful indicators of disruptive science and technology (Supplementary Tables S2-S4). While database errors in general do not only affect papers and patents with zero recorded references, they are especially problematic for these data entries, as having zero references causes a discontinuity in the CD index (Fig. 2**a,b**). Therefore, it is best practice to exclude zero reference papers and patents prior to further analysis. Indeed, many recent Science of Science publications [4, 7, 8] set the CD indices of papers that make zero references to non-defined.

We verified that our observations do not depend on the specific data source, the category within the respective data source, the choice of forward citation window, or the normalized CD indices (Extended Data Fig. A3, Supplementary Figs. S4-S8).

In summary, we revealed in three different ways that the decline in disruption, presented in Park et al. [1], is driven by papers and patents with zero references and $CD_5 = 1$. They remained hidden in the histograms, which the robustness checks failed to catch. Most of these papers and patents correspond to erroneous database entries. The curves showing how average CD indices have evolved, plotted in Park et al. [1], therefore, do not track declining disruption of scientific and technological work, but rather trace how metadata quality has increased over time.

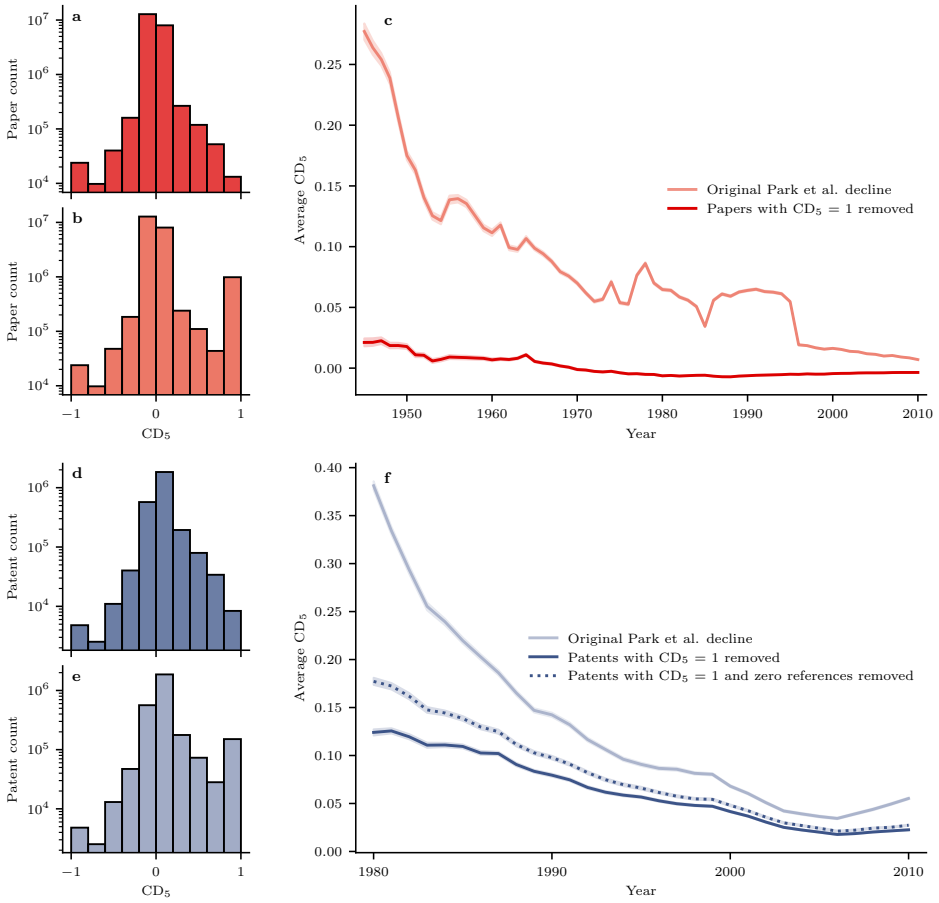

Fig. 1 | Distribution of the CD_5 index with vs without the hidden outliers and its impact on the apparent decline of disruptive science and technology. This figure shows that $CD_5 = 1$ papers and patents are driving the reported decline in the disruptiveness of scientific and technological knowledge over time for the *Web of Science* data source (with 22,479,429 papers) and the *PatentsView* data source (with 2,926,923 patents). For *PatentsView*, we also have access to sufficient metadata to exclude patents that make zero references, similarly impacting the decline. **a**, The distribution of the CD_5 index for papers in *Web of Science* as presented in Park et al. [1], created using the binwidth parameter in *seaborn 0.11.2*. This version of the library contains a bug regarding silently dropping the largest data points (1 in this case) when specifying the binwidth parameter [3]. **b,** The correct histogram for papers when using the bins parameter in *seaborn 0.11.2*. A peak at $CD_5 = 1$ is revealed with 972,161 additional papers. **c,** The time evolution of the average CD_5 index for papers. When dropping the hidden outliers with $CD_5 = 1$, the decline in disruptiveness almost completely disappears. The shaded bands correspond to 95% confidence intervals. Finally, note that the curve without $CD_5 = 1$ papers corresponds to (a), the histogram presented in Park et al. [1]. **d–f,** The equivalent plots for *PatentsView* revealing 142,362 additional patents with $CD_5 = 1$. When dropping the outliers with $CD_5 = 1$, the decline in disruptiveness reduces substantially. Unlike *Web of Science*, the *PatentsView* data source provided sufficient metadata to exclude patents with zero references, similarly impacting the data as removing outliers with $CD_5 = 1$ (Fig. 2 and Extended Data Fig. A2). Finally, note again that the curve without $CD_5 = 1$ patents corresponds to (d), the histogram presented in Park et al. [1].

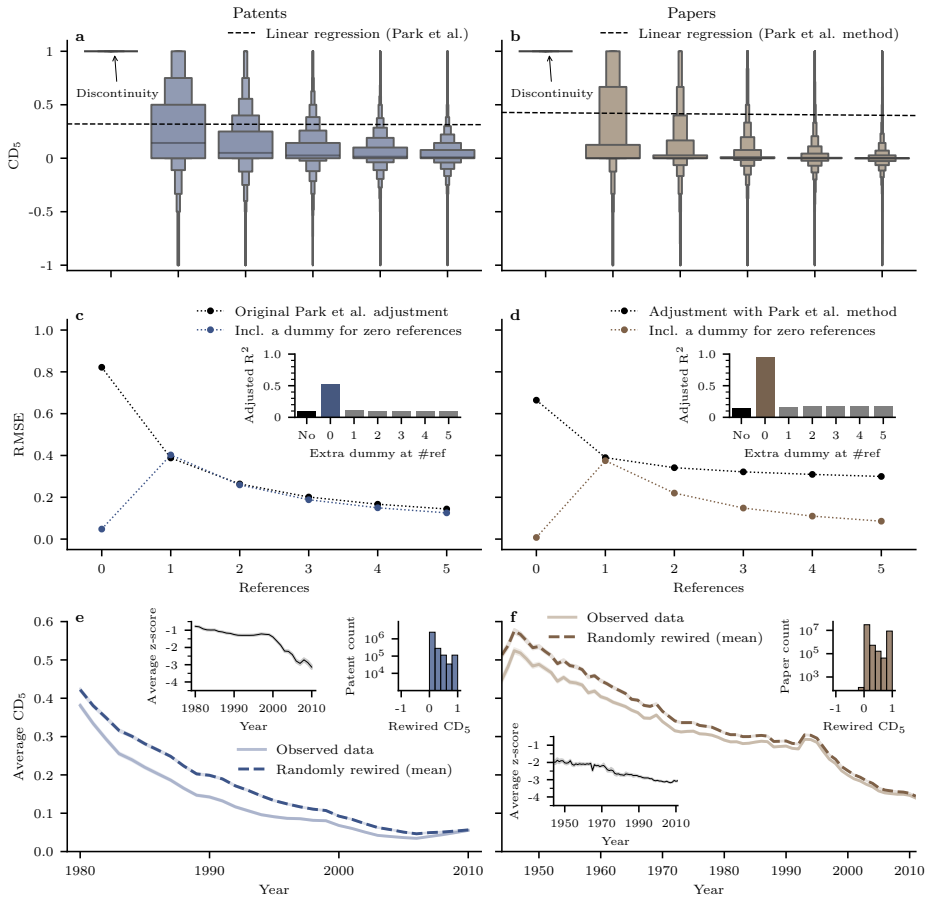

Fig. 2 | The reason why the robustness checks in Park et al. [1] failed to detect the consequences of the hidden outliers. This figure displays how the Park et al. [1] regression adjustment (models 4 and 8 in Supplementary Table 1 in [1]) fails to control for the discontinuous effect of zero references and that randomly rewired citation networks exhibit a similar temporal decline of CD_5. Results are shown for *PatentsView* (**a**, **c**, **e**; $n = 2,926,923$ patents) using the original Park et al. [1] data and *SciSciNet* [4] (**b**, **d**, **f**; $n = 39,888,199$ papers), replicating their *Web of Science* analysis. Shaded bands correspond to 95% confidence intervals. **a,** The distribution of the CD_5 per number of references is shown via letter-value plots which first identify the median, then extend boxes outward, each covering half of the remaining data [6]. Notably, in the case of zero references, the CD index is either one or remains undefined, causing a discontinuity. The marginal effect of references on CD_5 shows that the regression adjustment of Park et al. [1] fails to account for this discontinuity. **c,** The root mean squared errors (RMSE) show a pattern between the Park et al. [1] regression residuals and the number of references, showing that the model does not properly control for the discontinuous effect of zero references. Adding a dummy variable for zero references substantially improves the model fit as depicted by the adjusted R^2, while a similar effect is not found for other reference dummy variables. **e,** The average CD_5 of the rewired patent networks (mean over ten runs) mirrors the decline of the observed network over time. This close similarity is the result of the one-to-one correspondence between zero reference patents within the observed and simulated networks, as evidenced by the peak at one in the histogram of the rewired CD_5 shown in the inset plot. Finally, note that the gap between the observed CD_5 values and those from the simulated networks is becoming smaller over time, which implies that the decline in the z score found by Park et al. [1] and shown in the inset is the result of a decreasing standard deviation. **b**, **d**, **f**, The analogous, replicated plots for *SciSciNet*.

Acknowledgments. We thank M. Park, E. Leahey, and R.J. Funk for making their code and source material public, and for responding to a previous version of this report, highlighting the different robustness checks in their manuscript. We thank S.J. Klein (MIT) for helpful advice. We thank M. Waskom for maintaining the *seaborn* library open source which allowed us to quickly identify the bug in the plotting of the histograms.

Funding: Work was supported by the Research Council (OZR) of the VUB.
Competing interests: The authors declare no competing interests.
Ethics approval: Not applicable.
Consent to participate: Not applicable.
Consent for publication: Not applicable.
Availability of data and materials: The *Web of Science* and the *PatentsView* data for the study was retrieved from the public repository (https://doi.org/10.5281/zenodo.7258379) made publicly available by Park et al. [1]. The *SciSciNet* data source [4] (https://springernature.figshare.com/collections/SciSciNet_A_large-scale_open_data_lake_for_the_science_of_science_research/6076908/1) and the *DBLP-Citation-network V14* [9] (https://www.aminer.org/citation) are publicly available to download. The data for the reanalysis we performed on these two datasets are publicly available at https://github.com/VincentHolst/reanalysis_declining_disruption.
Code availability: The re-analysis code for this publication is publicly available at https://github.com/VincentHolst/reanalysis_declining_disruption. It is based on the original analysis code (https://doi.org/10.5281/zenodo.7258379) made publicly available by Park et al. [1] and uses the same software packages as Park et al. [1], that is *pandas 1.4.3*, *numpy 1.23.1*, *matplotlib 3.5.2* and *seaborn 0.11.2*. To replicate the regression table, we used StataMP v.18.0 (reghdfe v.6.12.3).
Authors' contributions: V.H. and V.G. were responsible for the main idea of the study. V.H. discovered the mistake in the original analysis (outliers hidden). A.A. discovered the reason why the mistake was made (seaborn update). V.H., A.A., and V.G. designed the analysis. A.A. replicated the regression analysis. V.H. and F.T. implemented the random rewiring algorithm. V.H. and F.T. tested the robustness of the results with additional data and analyses. V.H. and F.T. designed the final figures. S.W. independently reviewed the results. All authors discussed the results and collaboratively drafted and revised the manuscript.

Appendix A Extended Data

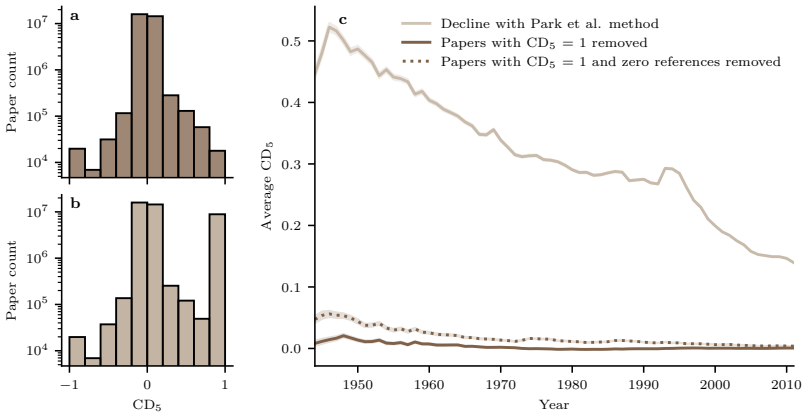

Fig. A1 | Distribution of the CD_5 index with vs without the hidden outliers and its impact on the disruptiveness for the *SciSciNet* data source. This figure replicates the observation that papers with $CD_5 = 1$ are driving the decline in disruptive science for the *SciSciNet* data source [4] (with 39,888,199 papers between 1944 and 2011), which originated from the *Microsoft Academic Graph*. **a**, The distribution of the CD_5 index for *SciSciNet*, created using the binwidth parameter in *seaborn 0.11.2*. Here again, the largest data points are hidden. **b,** The correct histogram of the underlying dataset. A peak at $CD_5 = 1$ is revealed, corresponding to 8,861,343 additional papers. **c,** The time evolution of the average CD_5 index. When dropping the outliers with $CD_5 = 1$, the decline in disruptiveness is negated. Excluding papers with zero references impacts the data similarly (Fig. 2 and Extended Data Fig. A2). The shaded bands correspond to 95% confidence intervals. Moreover, the curve with papers with $CD_5 = 1$ omitted is the curve corresponding to the histogram (a).

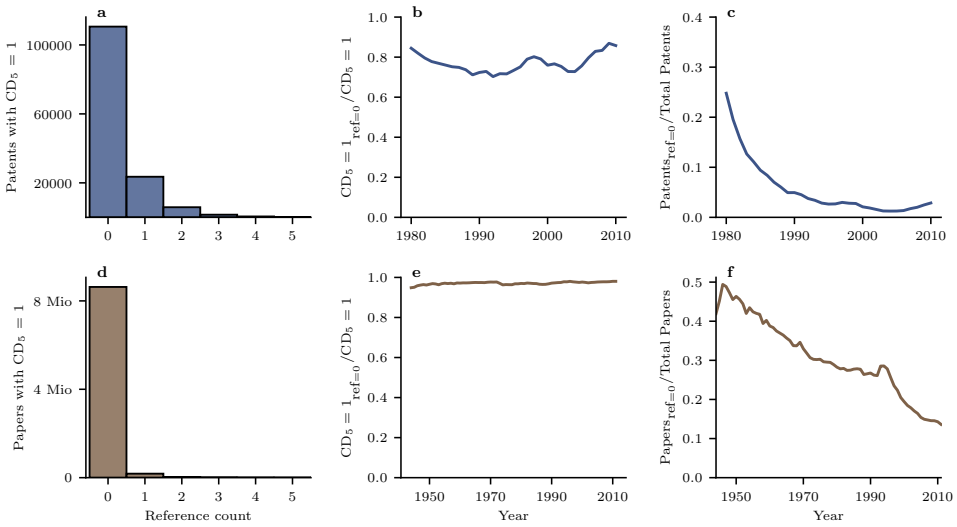

Fig. A2 | Papers and patents with $CD_5 = 1$ predominantly make zero references. This figure displays that most papers in the *SciSciNet* data source [4] ($n = 39,888,199$) and most patents in the *PatentsView* data source ($n = 2,926,923$) with $CD_5 = 1$ have zero references. **a,** Our analysis shows that *PatentsView* contains $142,362$ patents with $CD_5 = 1$ between 1980 and 2010, of which 78 % appear in the database with zero references. **b,** Within the category of patents with $CD_5 = 1$, the relative frequency of patents with zero references is stable between 1980 and 2010. **c,** The relative frequency of patents with CD_5 index exactly equal to one and zero references is decreasing over time. Therefore, a substantial part of the reported decline in the disruptiveness of technological knowledge over time can be attributed to a relatively increasing metadata quality over time. It is also intriguing to note how well the shape of this curve resembles the shape of the top curve shown in Fig. 1**f**. **d,** *SciSciNet* [4] shows a similar behaviour with $8,861,343$ papers having $CD_5 = 1$ between 1944 and 2011, of which 97 % appear in the database with zero references. **e,** Within the category of papers with $CD_5 = 1$, the relative frequency of papers with zero references is stable between 1944 and 2011. **f,** The relative frequency of papers with CD_5 index exactly equal to one and zero references is decreasing over time. Therefore, a substantial part of the observed decline in the disruptiveness of scientific knowledge over time can be attributed to a relatively increasing metadata quality over time. It is also intriguing to note how well the shape of this curve resembles the shape of the top curve shown in Extended Data Fig. A1**c**.

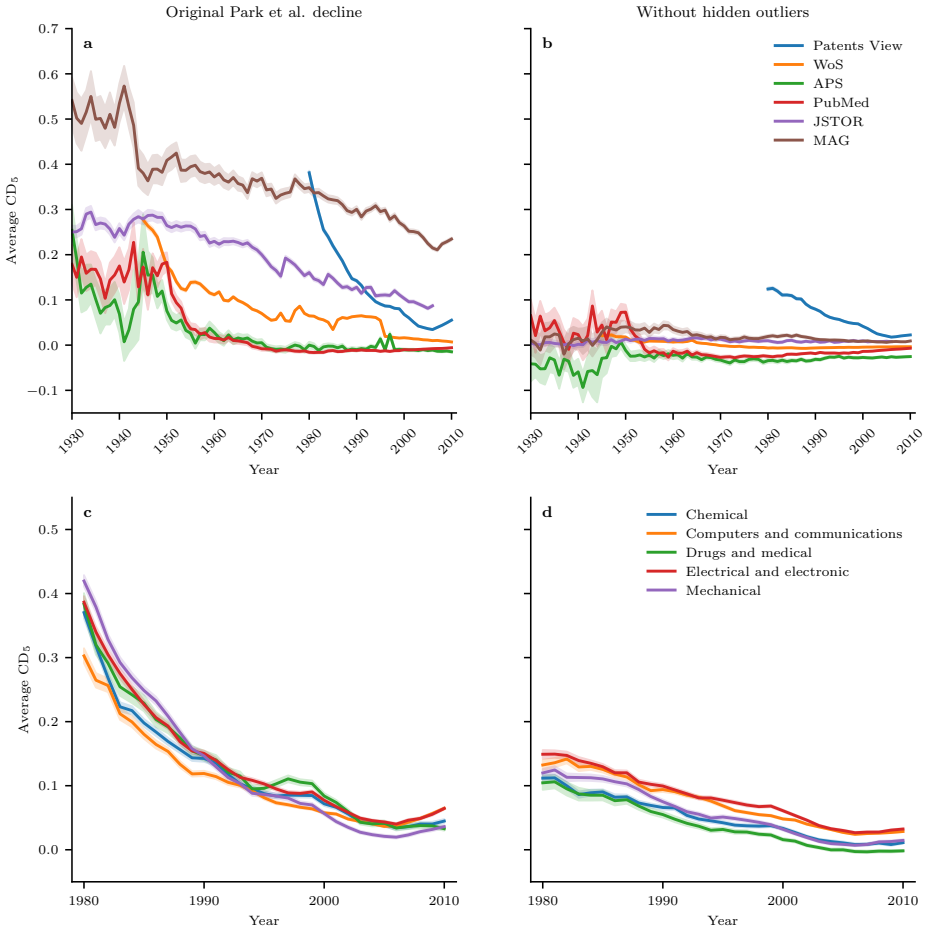

Fig. A3 | Across various data sources and within different categories, papers and patents with $CD_5 = 1$ are driving the decline in the disruptiveness in scientific and technological knowledge over time. This figure displays the average CD_5 index over time for six data sources and five different patent categories. The data sources are *JSTOR* (1,588,088 papers), the *American Physical Society* corpus (461,359 papers), *Microsoft Academic Graph* (random sample of 1,000,000 papers), and *PubMed* (1,563,211 papers). For reference, the *Web of Science* (22,479,429 papers) and *PatentsView* (2,926,923 patents) data sources are also included. The patent categories are *Chemical* (517,964 patents), *Computers and communications* (748,849 patents), *Drugs and medical* (321,449 patents), *Electrical and electronic* (734,769 patents) and *Mechanical* (603,892 patents). Shaded bands correspond to 95% confidence intervals. **a**, The temporal evolution of the average CD_5 index for different data sources as presented in Park et al. [1] (Extended Data Fig. 6 in [1]). **b**, The time evolution of the average CD_5 index for different data sources after removing the outliers with $CD_5 = 1$ from the data sources. For all mentioned data sources that encompass papers, the decline in the disruptiveness almost completely disappears. For the *PatentsView* data source, the decline in the disruptiveness also reduces notably. **c**, The time evolution of the average CD_5 index for different patent categories as presented in Park et al. [1] (Fig. 2**b** in [1]). **d**, The time evolution of the average CD_5 index for different patent categories after removing the outliers with $CD_5 = 1$ from the categories. We see that the decline in disruptiveness reduces similarly across all five categories.

SI Guide for "Dataset Artefacts are the Hidden Drivers of the Declining Disruptiveness in Science"

Vincent Holst, Data Analytics Laboratory, Vrije Universiteit Brussel
Andres Algaba, Data Analytics Laboratory, Vrije Universiteit Brussel
Floriano Tori, Data Analytics Laboratory, Vrije Universiteit Brussel
Sylvia Wenmackers, Centre for Logic and Philosophy of Science (CLPS), KU Leuven
Vincent Ginis, Data Analytics Laboratory, Vrije Universiteit Brussel & School of Engineering and Applied Sciences, Harvard University

Supplementary Methods

Here, we provide additional mathematical justifications, supplementary tables and supplementary figures for our reanalysis of Park et al. [1].

Table of Contents

Supplementary Information

S1. The CD index

Let $\mathcal{G} = (V, E)$ be a directed citation network. The set V corresponds to the papers in A and E corresponds to the citations between the papers. The adjacency matrix $A = (a_{ij})$ of $\mathcal{G} = (V, E)$ is given by $a_{ij} = 1$ if and only if paper i cites paper j, and $a_{ij} = 0$ otherwise. Every paper $i \in V$ is assigned a publication date d_i, usually given in datetime format. The directed citation network $\mathcal{G}$ possesses a temporal structure: if paper i cites paper j, we have $d_i > d_j$, i.e. paper i was published after paper j.

Let $i \in V$ be a focal paper with publication date d_i and $t \in \mathbb{N}^+ = \{1, 2, 3, ...\}$ a forward citation window. Let $U \subset V$ be the papers published between $(d_i, d_i + t$ years], e.g., if d_i is equal to 1984-01-01, then U encompasses all papers published after d_i until 1989-01-01. We consider the following sets:

$$\begin{aligned} F &:= \{j \in U |\, j \text{ cites the focal paper } i \text{ but none of its references}\}, \\ B &:= \{j \in U |\, j \text{ cites both the focal paper } i \text{ and at least one of its references}\}, \\ R &:= \{j \in U |\, j \text{ does not cite the focal paper } i \text{ but at least one of its references}\}. \end{aligned} \tag{S1}$$

Let $N_F = |F|$, $N_B = |B|$ and $N_R = |R|$. Then the CD_t index [2] of paper i is given by:

$$\mathrm{CD}_t = \frac{N_F - N_B}{N_F + N_B + N_R}. \tag{S2}$$

If paper i has zero references to prior work, the sets B and R are empty by default and it is easy to see that $\mathrm{CD}_t = (N_F - N_B)/(N_F + N_B + N_R) = N_F/N_F$ is either exactly equal to one (if F is not empty, i.e. if i receives at least one forward citation within t years after publication) or remains undefined (if F is empty, i.e. if i receives no forward citation within t years after publication).

The *DBLP-Citation-network V14* [9] provides the publication date only in YYYY format. If the focal paper i is published in a given year d_i, we considered the subsets $U \subset V$ of papers published between $[d_i + 1, d_i + 5]$ and $W \subset V$ of papers published between $[d_i, d_i + 5]$ as follows in the calculation of the CD_5 index:

$$\begin{aligned} F &:= \{j \in W |\, j \text{ cites the focal paper } i \text{ but none of its references}\}, \\ B &:= \{j \in W |\, j \text{ cites both the focal paper } i \text{ and at least one of its references}\}, \\ R &:= \{j \in U |\, j \text{ does not cite the focal paper } i \text{ but at least one of its references}\}. \end{aligned} \tag{S3}$$

S2. Regression adjustment

Park et al. [1] use a linear regression to control for potential changes in citation patterns by including the number of references on the paper/patent level, together with fixed effects and additional control variables at the (sub-)field and year level. We also include a zero references dummy variable, which is equal to one if the paper/patent has zero references and zero else, to explicitly control for the discontinuous effect of zero references (Fig. 2**a,b**). The regression looks as follows:

$$\begin{aligned}\mathrm{CD5}_{i,t(i),k(i)} = \alpha + \underbrace{\sum_{t=1}^{T-1} \theta_t \ \mathrm{year}_{t(i)}}_{\substack{\text{time}\\ \text{fixed effects}}} + \underbrace{\sum_{k=1}^{K-1} \delta_k \ \text{(sub-)field}_{k(i)}}_{\substack{\text{(sub-)field}\\ \text{fixed effects}}} + \underbrace{\beta_1 \ \#\text{references}_i}_{\substack{\text{paper/patent}\\ \text{level control}}} \\ + \underbrace{\gamma_1 \ \#\text{patents/papers}_{t(i),k(i)} + \gamma_2 \ \overline{\#\text{references}}_{t(i),k(i)} + \gamma_3 \ \overline{\#\text{authors/inventors}}_{t(i),k(i)}}_{\substack{\text{(sub-)field and year}\\ \text{level controls}}} \\ + \underbrace{\zeta \ \text{zero references}_i}_{\substack{\text{zero references}\\ \text{dummy variable}}} + \varepsilon_i, \end{aligned} \tag{S4}$$

where i, t, and k denote the paper/patent, the publication/grant year, and (sub-)field, respectively. Moreover, $t(i)$ and $k(i)$ indicate that the publication/grant year t and (sub-)field k depend on the paper/patent i, # denotes "the number of," and $\overline{x}$ is the average of a variable x. The time and (sub-)field fixed effects can be estimated by including dummy variables for the publication/grant years and (sub-)fields, which are equal to one if the year or field is equal to the year or field of the current paper/patent and zero else, with the exclusion of a reference category (i.e., $T-1$ and $K-1$). Note that we include the zero references dummy variable to explicitly control for the discontinuous effect and to show the consequences for the main findings of Park et al. [1] by displaying the regression adjusted CD_5 (Supplementary Fig. S1 and Supplementary Table S1). Since we only want to show that the regression model of Park et al. [1] does not control for the discontinuity of zero references, we do not make any further changes to the regression model, despite other potential improvements, such as taking the natural logarithm of the number of references as a control variable [10].

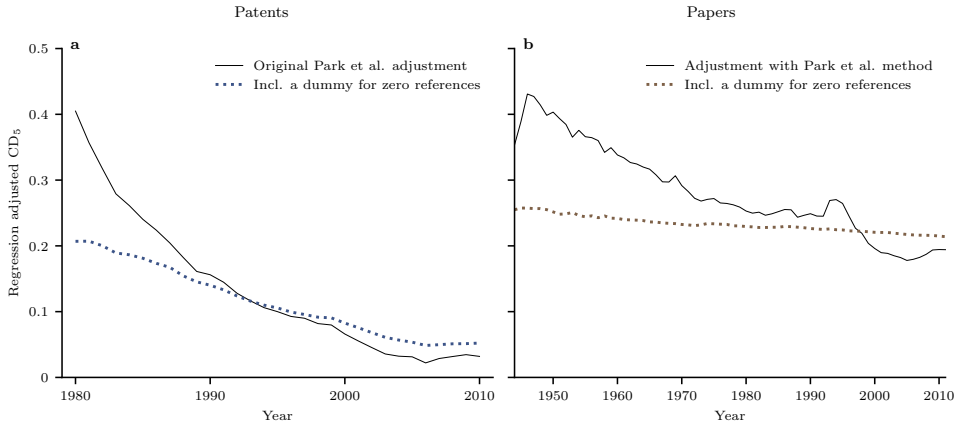

Fig. S1 | Inclusion of the zero references dummy variable to properly control for the corresponding discontinuity in the CD_5 largely negates the decline of disruptive science. This figure displays that the inclusion of the zero references dummy variable to account for the corresponding discontinuity in the CD_5 (Fig. 2**a,b**) in the regression adjustment substantially reduces the decline for the *PatentsView* data source (with $2,926,923$ patents, models (3) and (4) in Supplementary Table S1) and largely negates the decline for the *SciSciNet* data source [4] (with $39,888,199$ papers, models (1) and (2) in Supplementary Table S1). **a,** As displayed in Fig. 2**a,c**, the linear regression conducted by Park et al. [1] for *PatentsView* (model 8 from Supplementary Table 1 in [1]) fails to control for the discontinuous effect of zero references. This is shown by including a dummy variable explicitly controlling for zero references, resulting in a substantial increase in the adjusted R^2, an effect not observed for any other number of references. Here, we show that explicitly controlling for the discontinuous effect of zero references substantially reduces the temporal decline of the regression adjusted CD_5. **b,** The analogous analysis for *SciSciNet*. Here, we replicate the Park et al. [1] regression model conducted for *Web of Science* (model 4 from Supplementary Table 1 in [1]). Notably, explicitly controlling for the discontinuous effect of zero references largely negates the temporal decline of the regression adjusted CD_5.

Table S1 Regression models for adjusting CD_5.

	SciSciNet		PatentsView	
Variables	(1)	(2)	(3)	(4)
Year=1945	0.04^{***}	0.00		
Year=1946	0.08^{***}	0.00		
Year=1947	0.07^{***}	0.00		
Year=1948	0.06^{***}	0.00		
Year=1949	0.05^{***}	0.00		
Year=1950	0.05^{***}	-0.00^{*}		
Year=1951	0.04^{***}	-0.01^{***}		
Year=1952	0.03^{***}	-0.01^{***}		
Year=1953	0.01^{***}	-0.00^{**}		
Year=1954	0.02^{***}	-0.01^{***}		
Year=1955	0.01^{***}	-0.01^{***}		
Year=1956	0.01^{**}	-0.01^{***}		
Year=1957	0.01	-0.01^{***}		
Year=1958	-0.01^{***}	-0.01^{***}		
Year=1959	-0.00	-0.01^{***}		
Year=1960	-0.02^{***}	-0.01^{***}		
Year=1961	-0.02^{***}	-0.02^{***}		
Year=1962	-0.03^{***}	-0.02^{***}		
Year=1963	-0.03^{***}	-0.02^{***}		
Year=1964	-0.03^{***}	-0.02^{***}		
Year=1965	-0.04^{***}	-0.02^{***}		
Year=1966	-0.05^{***}	-0.02^{***}		
Year=1967	-0.06^{***}	-0.02^{***}		
Year=1968	-0.06^{***}	-0.02^{***}		
Year=1969	-0.05^{***}	-0.02^{***}		
Year=1970	-0.06^{***}	-0.02^{***}		
Year=1971	-0.07^{***}	-0.02^{***}		
Year=1972	-0.08^{***}	-0.02^{***}		
Year=1973	-0.09^{***}	-0.02^{***}		
Year=1974	-0.08^{***}	-0.02^{***}		
Year=1975	-0.08^{***}	-0.02^{***}		
Year=1976	-0.09^{***}	-0.02^{***}		
Year=1977	-0.09^{***}	-0.02^{***}		
Year=1978	-0.09^{***}	-0.02^{***}		
Year=1979	-0.09^{***}	-0.03^{***}		
Year=1980	-0.10^{***}	-0.03^{***}		
Year=1981	-0.10^{***}	-0.03^{***}	-0.05^{***}	0.00
Year=1982	-0.10^{***}	-0.03^{***}	-0.09^{***}	-0.01^{***}
Year=1983	-0.11^{***}	-0.03^{***}	-0.13^{***}	-0.02^{***}
Year=1984	-0.11^{***}	-0.03^{***}	-0.14^{***}	-0.02^{***}
Year=1985	-0.10^{***}	-0.03^{***}	-0.16^{***}	-0.03^{***}
Year=1986	-0.10^{***}	-0.03^{***}	-0.18^{***}	-0.03^{***}
Year=1987	-0.10^{***}	-0.03^{***}	-0.20^{***}	-0.04^{***}
Year=1988	-0.11^{***}	-0.03^{***}	-0.22^{***}	-0.05^{***}
Year=1989	-0.11^{***}	-0.03^{***}	-0.24^{***}	-0.06^{***}
Year=1990	-0.10^{***}	-0.03^{***}	-0.25^{***}	-0.07^{***}
Year=1991	-0.11^{***}	-0.03^{***}	-0.26^{***}	-0.07^{***}
Year=1992	-0.11^{***}	-0.03^{***}	-0.28^{***}	-0.08^{***}
Year=1993	-0.09^{***}	-0.03^{***}	-0.29^{***}	-0.09^{***}
Year=1994	-0.08^{***}	-0.03^{***}	-0.30^{***}	-0.10^{***}
Year=1995	-0.09^{***}	-0.03^{***}	-0.30^{***}	-0.10^{***}
Year=1996	-0.11^{***}	-0.03^{***}	-0.31^{***}	-0.11^{***}
Year=1997	-0.13^{***}	-0.03^{***}	-0.32^{***}	-0.11^{***}
Year=1998	-0.13^{***}	-0.03^{***}	-0.32^{***}	-0.12^{***}
Year=1999	-0.15^{***}	-0.03^{***}	-0.33^{***}	-0.12^{***}
Year=2000	-0.16^{***}	-0.03^{***}	-0.34^{***}	-0.12^{***}
Year=2001	-0.16^{***}	-0.04^{***}	-0.35^{***}	-0.13^{***}
Year=2002	-0.17^{***}	-0.04^{***}	-0.36^{***}	-0.14^{***}
Year=2003	-0.17^{***}	-0.04^{***}	-0.37^{***}	-0.15^{***}
Year=2004	-0.17^{***}	-0.04^{***}	-0.37^{***}	-0.15^{***}
Year=2005	-0.18^{***}	-0.04^{***}	-0.37^{***}	-0.15^{***}
Year=2006	-0.17^{***}	-0.04^{***}	-0.38^{***}	-0.16^{***}
Year=2007	-0.17^{***}	-0.04^{***}	-0.38^{***}	-0.16^{***}
Year=2008	-0.17^{***}	-0.04^{***}	-0.37^{***}	-0.16^{***}
Year=2009	-0.16^{***}	-0.04^{***}	-0.37^{***}	-0.16^{***}
Year=2010	-0.16^{***}	-0.04^{***}	-0.37^{***}	-0.16^{***}
Year=2011	-0.16^{***}	-0.04^{***}		
β_1	$-4.73e{-3}^{***}$	$-2.73e{-4}^{***}$	$-1.11e{-3}^{***}$	$-6.46e{-4}^{***}$
γ_1	$-3.20e{-7}^{***}$	$1.23e{-8}^{***}$	$3.17e{-6}^{***}$	$5.67e{-7}^{***}$
γ_2	$5.94e{-4}^{***}$	$1.03e{-4}^{***}$	$9.61e{-4}^{***}$	$7.78e{-4}^{***}$
γ_3	$1.23e{-2}^{***}$	$3.93e{-3}^{***}$	$2.64e{-2}^{***}$	$1.77e{-2}^{***}$
ζ		0.98^{***}		0.90^{***}
α	0.43^{***}	0.03^{***}	0.32^{***}	0.13^{***}
(Sub-)field fixed effects	yes	yes	yes	yes
N	39,888,199	39,888,199	2,926,923	2,926,923
Adjusted R^2	0.15	0.95	0.10	0.52

Note: Model 3 is Model 8 from [1] (Suppl. Table 1). Model 1 replicates Model 4 from [1] (Suppl. Table 1) on SciSciNet instead of WoS. Models 2 and 4 control for zero references by including a dummy variable. Estimates are from an OLS-regression (Eq. S4) and significance levels are for a two-sided t-test with a H_0 of the regression coefficient being equal to zero ($^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$).

S3. Monte Carlo simulations

In their original manuscript, Park et al. [1] conducted Monte Carlo simulations to check if the observed decline of CD_5 is caused by changes in the citation networks' general topology instead of societal processes. Therefore, Park et al. [1] used a random rewiring algorithm [5] that preserves the topological structure, i.e. the in- and outdegree (resp. number of forward citations and references) of the involved papers and patents, and the age structure, i.e. the publications years of the involved papers and patents, but randomly rewires the citations between the involved papers and patents.

Park et al. [1] described the rewiring algorithm as follows: if paper A cites paper B and paper C cites paper D, then the switch to paper A cites paper D and paper C cites paper B is retained if and only if (1) paper A and paper C (resp. paper B and paper D) have the same number of forward citations (resp. references) after the switch and (2) paper A and paper C (resp. paper B and paper D) were published in the same year. The random rewiring is repeated until $100 \cdot \#\text{edges}$ switches are retained.

To evaluate the Monte Carlo simulations, Park et al. [1] calculated an average z score among papers or patents published in each year. For an individual paper or patent, the z score is given by:

$$\frac{\mathrm{CD}_{\mathrm{observed}} - \mu_{\mathrm{rewired}}}{\sigma_{\mathrm{rewired}}}. \tag{S5}$$

Here, $\mathrm{CD}_{\mathrm{observed}}$ denotes the observed CD_5 index in the unaltered data source, μ_{rewired} denotes the average CD_5 index of the same paper or patent calculated across ten randomly rewired citation networks and $\sigma_{\mathrm{rewired}}$ denotes the corresponding standard deviation. Based on the temporal decline of the average z score (Extended Data Fig. 8**c,f** in [1]) the authors concluded "We find that on average, papers and patents tend to be less disruptive than would be expected by chance, and moreover, the gap between the observed CD index values and those from the randomly rewired networks is increasing over time, which is consistent with our findings of a decline in disruptive science and technology." However, the main findings in Park et al. [1] are based on the decline of the average CD_5 over time (Fig. 2 in [1]). To test the robustness of these results against random rewiring, it is therefore logical to also evaluate the rewired CD_5 against time (Fig. 2**e,f**, Supplementary Fig. S2, S3, S5). These plots unambiguously show that the aforementioned gap is, in fact, narrowing over time. The temporal decrease of the average z score can therefore be attributed to the following phenomenon: the gap between the rewired and observed CD indices shown in Fig. 2**e,f** corresponds to the (averaged) numerator of the above equation, which indicates that the (averaged) denominator of the above equation, the variance/standard deviation within the ten randomly rewired citation networks, decreases over time (with the caveat that the mean of $\frac{a}{b}$ is of course not exactly equal to the mean of a divided by the mean of b).

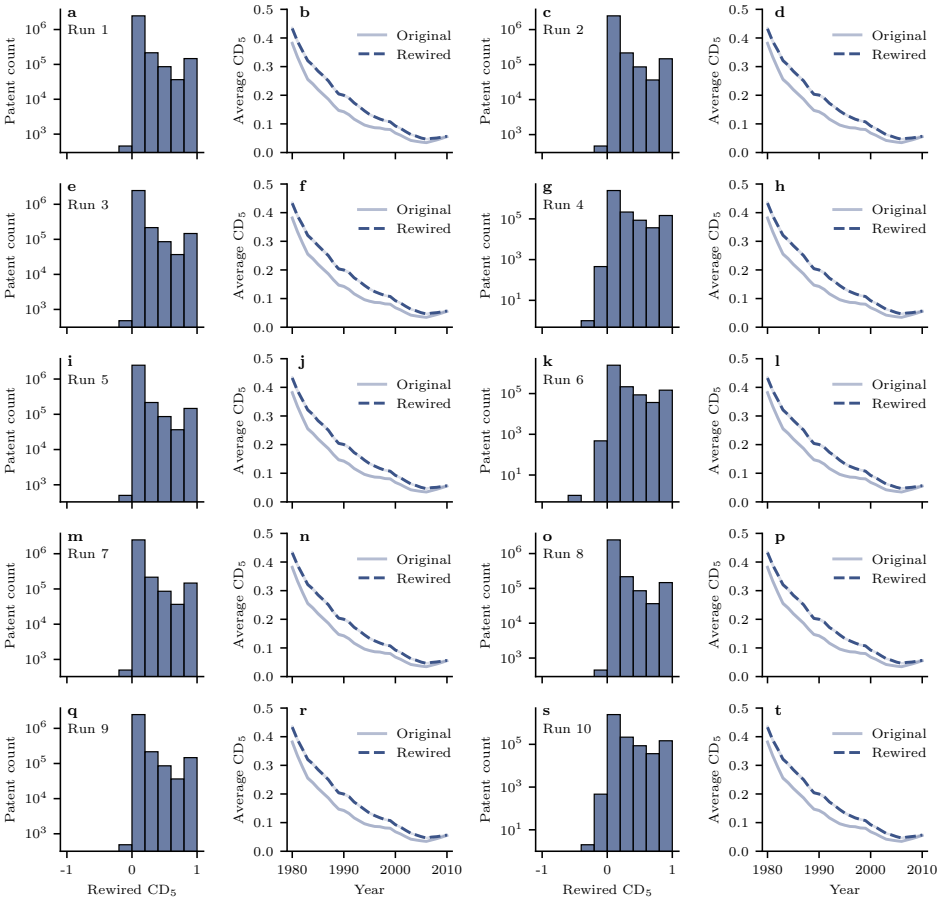

Fig. S2 | The temporal decline of the CD_5 index for patents is mirrored by random citation behaviour supporting that the hidden outliers are driving the decline. This figure displays the distribution (panels **a**, **c**, **e**, **g**, **i**, **k**, **m**, **o**, **q**, **s**) and time average (panels **b**, **d**, **f**, **h**, **j**, **l**, **n**, **p**, **r**, **t**) of the CD_5 index for the ten randomly rewired *PatentsView* data sources (with $2,926,923$ patents) from [1]. Here, the random rewiring algorithm [5] used by Park et al. [1] preserves the in- and outdegree (resp. forward citations and references) and publication year of the involved patents. In particular, this induces a one-to-one correspondence between the zero reference patents in the rewired and original network. The shaded bands in the plots correspond to 95 % confidence intervals. **a**, The distribution of the rewired CD_5 for the first of the ten randomly rewired patent networks shows that the algorithm used by Park et al. [1] boosts CD index values. This is unsurprising, as the CD index measures triadic closure and randomly rewiring a sparse citation network naturally reduces the number of triangles. Also, note that the histogram still shows the peak at one, confirming the aforementioned one-to-one correspondence. **b**, For the first of the ten rewired patent networks, the temporal decline of the rewired CD_5 mirrors the decline of the original patent network. Since the zero reference patents are preserved by the rewiring algorithm and the majority of the hidden outliers make zero references (Extended Data Fig. A2**a**), this observation provides yet another proof that the hidden outliers are driving the decline. In other words, even upon random citation behaviour, having a certain relative number of zero reference patents with $CD_5 = 1$ per year induces a decline nearly identical to the one reported in the original manuscript of Park et al. [1] (Extended Data Fig. A2**c**). **c**–**t**, The equivalent plots for the remaining nine rewiring runs show similar results.

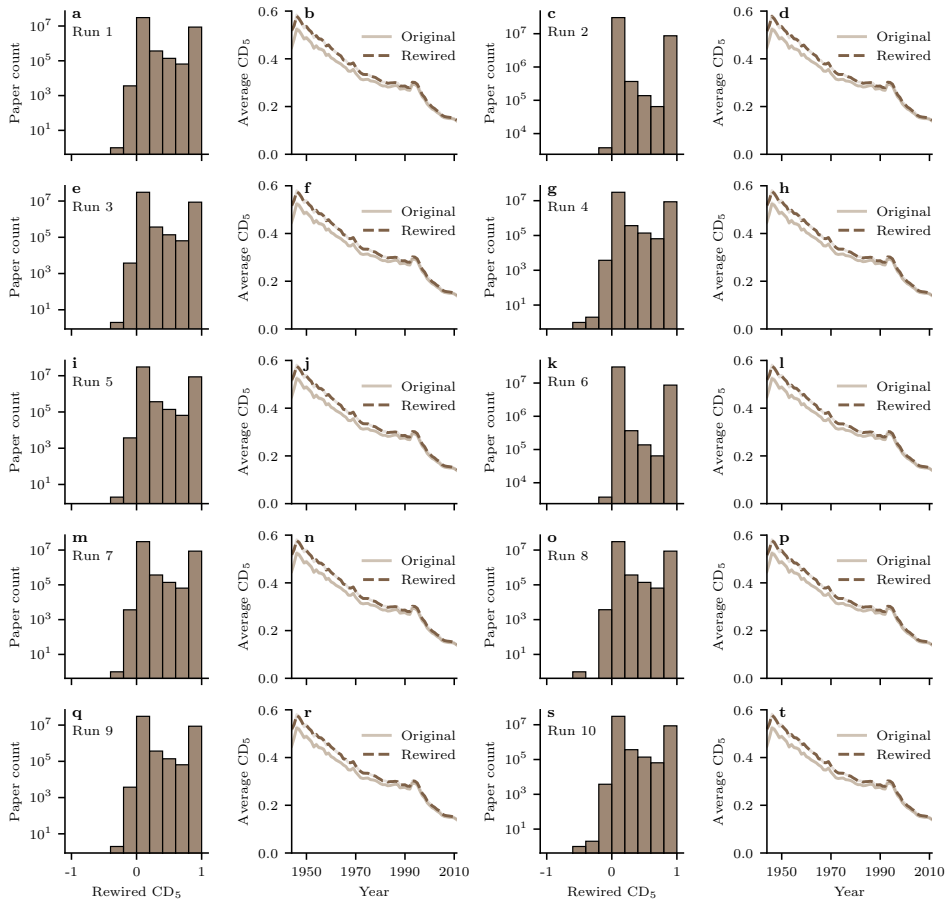

Fig. S3 | The temporal decline of the CD_5 index for the *SciSciNet* data source is mirrored by random citation behaviour supporting that the hidden outliers are driving the decline. This figure displays the distribution (panels **a**, **c**, **e**, **g**, **i**, **k**, **m**, **o**, **q**, **s**) and time average (panels **b**, **d**, **f**, **h**, **j**, **l**, **n**, **p**, **r**, **t**) of the CD_5 index for ten randomly rewired *SciSciNet* data sources [4] (with 39,888,199 papers), replicating the findings of Supplementary Fig. S2. We use the same random rewiring algorithm [5] as Park et al. [1]. The shaded bands in the plots correspond to 95 % confidence intervals. **a**, The distribution of the rewired CD_5 for the first of the ten randomly rewired paper networks shows that the algorithm used by Park et al. [1] boosts CD index values. The peak at one indicates that random rewiring preserves the zero reference papers. **b**, For the first of the ten rewired papers networks, the temporal decline of the rewired CD_5 mirrors the decline of the original papers network. Since the zero reference papers are preserved by the rewiring algorithm and the majority of the hidden outliers make zero references (Extended Data Fig. A2**d**), this observation provides yet another proof that the hidden outliers are driving the decline. **c**–**t**, The equivalent plots for the remaining nine rewiring runs show similar results.

S4. DBLP citation network

In this section, we replicate the same analysis for another independent paper dataset, the *DBLP-Citation-network V14* [9]. This data source contains $1,683,086$ papers published between 1970 and 2010 in the field of Computer Science.

- Fig. S4 shows analogous observations for the *DBLP-Citation-network V14* datasets as Fig. 1 by replicating the CD_5 with and without outliers. Since the *DBLP-Citation-network V14* only contains the publication dates of the papers in YYYY format, the CD_5 index is calculated as described in Supplementary Equation S3.

- Fig. S5 performs the rewiring analysis for the *DBLP-Citation-network V14*.

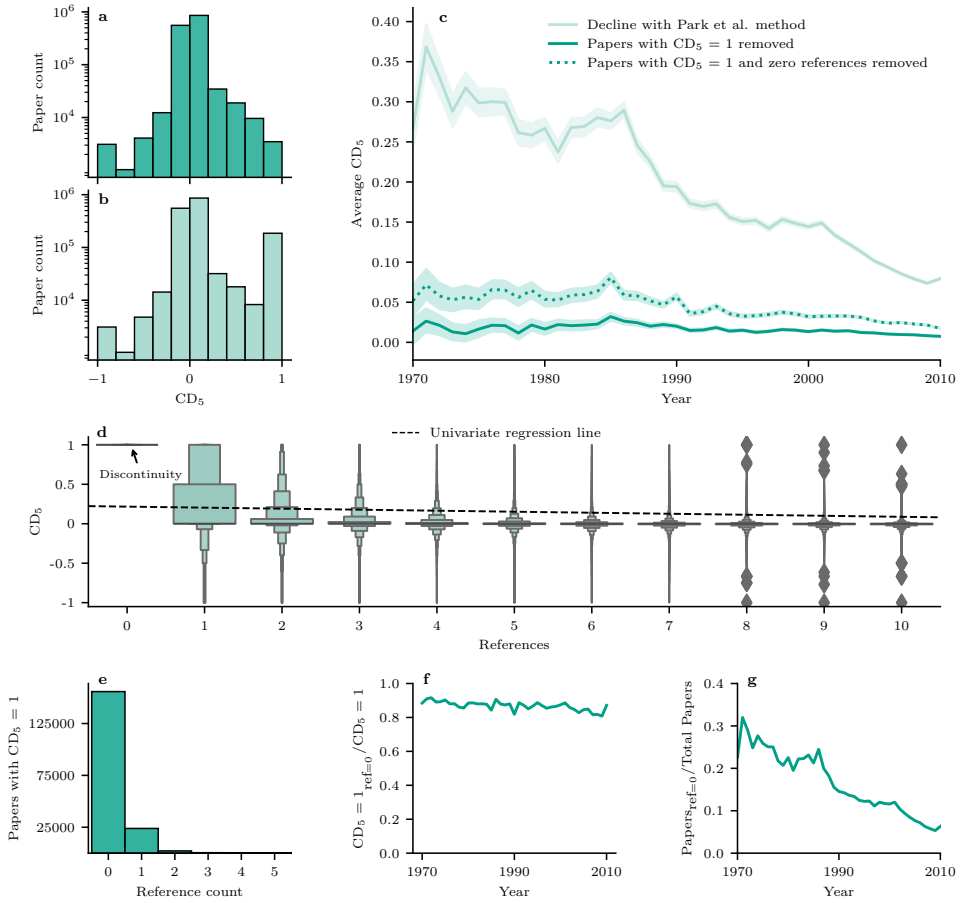

Fig. S4 | Distribution of the CD_5 index with vs without the hidden outliers and its impact on the disruptiveness for the *DBLP-Citation-network V14*. This figure replicates the observation that papers with $CD_5 = 1$ are driving the decline in disruptive science the *DBLP-Citation-network V14* [9]. **a**, The distribution of the CD_5 index, created using the binwidth parameter in *seaborn 0.11.2*. Here again, the largest data points are hidden. **b,** The correct histogram of the underlying dataset. A peak at $CD_5 = 1$ is revealed, corresponding to 182,398 additional papers. **c,** The time evolution of the average CD_5 index. When dropping the outliers with $CD_5 = 1$, the decline in disruptiveness is negated. Removing papers with zero references impacts the decline similarly. Moreover, the curve with papers with $CD_5 = 1$ omitted is the curve corresponding to the histogram (a). The shaded bands correspond to 95% confidence intervals. **d,** The distribution of the CD_5 per number of references is shown via letter-value plots which first identify the median, then extend boxes outward, each covering half of the remaining data [6]. The univariate regression line shows that an ordinary least squared regression fails to capture the discontinuous effect of zero references (Fig. 2 **a,b**). **e,** The *DBLP-Citation-network V14* contains 182,398 papers with $CD_5 = 1$ between 1970 and 2010, of which 85 % appear in the database with zero references. **f,** Within the category of papers with $CD_5 = 1$, the relative frequency of papers with zero references is stable between 1970 and 2010. **g,** The relative frequency of papers with CD_5 index exactly equal to one and zero references is decreasing over time, resembling the shape of the top curve shown in panel (c).

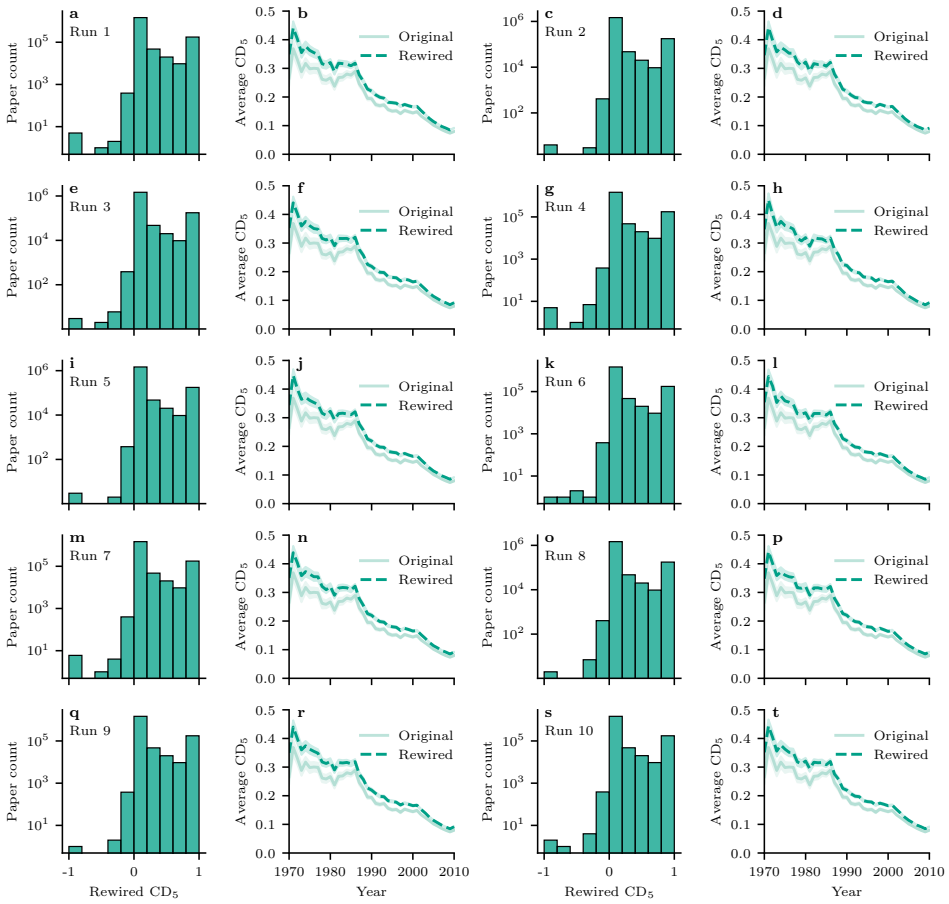

Fig. S5 | The temporal decline of the CD_5 index for the *DBLP-Citation-network V14* is mirrored by random citation behaviour supporting that the hidden outliers are driving the decline. This figure displays the distribution (panels **a**, **c**, **e**, **g**, **i**, **k**, **m**, **o**, **q**, **s**) and time average (panels **b**, **d**, **f**, **h**, **j**, **l**, **n**, **p**, **r**, **t**) of the CD_5 index for ten randomly rewired *DBLP-Citation-network V14* data sources, replicating Supplementary Fig. S2 and S3. We use the same random rewiring algorithm [5] as Park et al. [1]. The shaded bands in the plots correspond to 95% confidence intervals. **a**, The distribution of the rewired CD_5 for the first of the ten randomly rewired paper networks shows that the algorithm used by Park et al. [1] boosts CD index values. The peak at one indicates that random rewiring preserves the zero reference papers. **b**, For the first of the ten rewired papers networks, the temporal decline of the rewired CD_5 mirrors the decline of the original network. Since the zero reference papers are preserved by the rewiring algorithm and the majority of the hidden outliers make zero references (Supplementary Fig. S4**e**), this observation provides yet another proof that the hidden outliers are driving the decline. **c**–**t**, The equivalent plots for the remaining nine rewiring runs show similar results.

S5. Different forward citation windows

In this section, we analyse the role of the outliers in the decline for CD indices with different forward citation windows. We do this for both the *SciSciNet* [4] and the *PatentsView* data source.

- Fig. S6 shows the same analysis as Fig. 1 with the CD_{10} index for both *SciSciNet* and *PatentsView*. Contrary to the CD_5 index, the CD_{10} index considers forward citations published within 10 years after the publication of the focal paper.

- Fig. S7 shows the same analysis as Fig. 1 with the CD_{max} index for both *SciSciNet* and *PatentsView*. Contrary to the CD_5 index, the CD_{max} index considers all forward citations of a focal paper or patent. For *SciSciNet*, we used the precomputed disruption indices provided by [4] for papers with at least one forward citation and one reference. To allow comparison with the Park et al. [1] method, we imputed the values with at least one forward citation and zero references to one and the values with zero forward citations and at least one reference to zero.

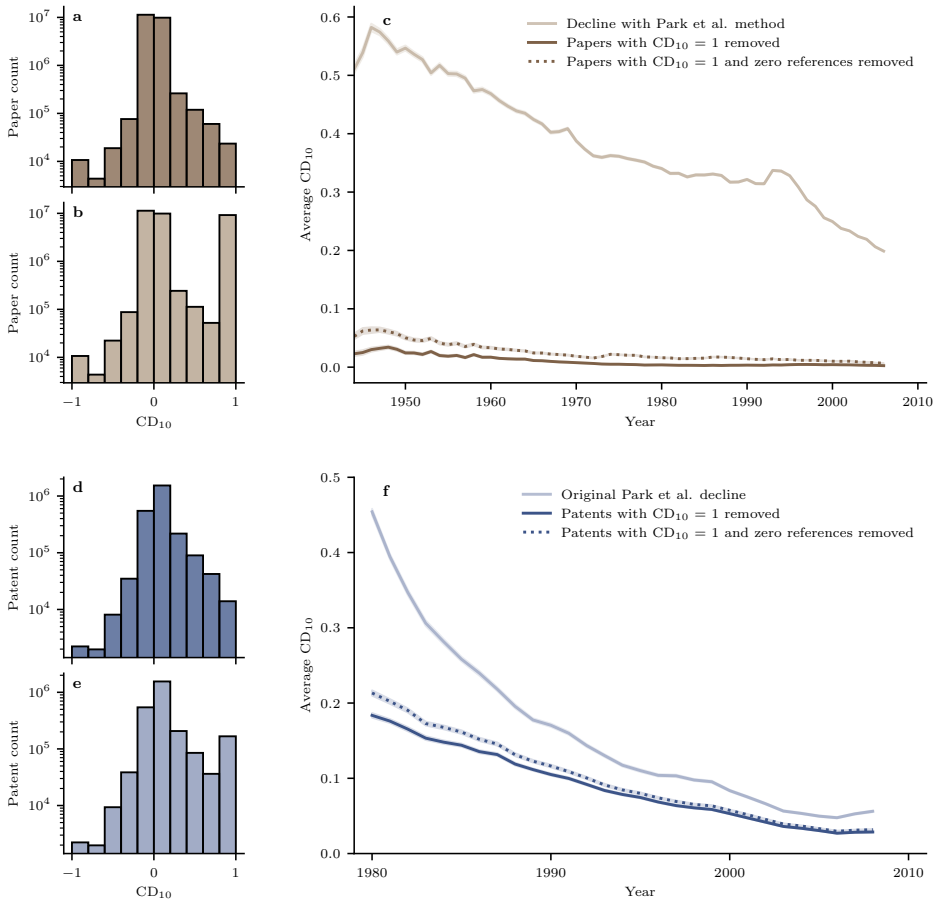

Fig. S6 | Distribution of the CD_{10} index with vs without the hidden outliers and its impact on the disruptiveness for the *SciSciNet* and the *PatentsView* data source. This figure displays the distribution and time average of the CD_{10} index (computed over a forward citation window of ten years) for the *SciSciNet* [4] data source (with $30,982,865$ papers until 2006) and the *PatentsView* data source (with $2,645,344$ patents until 2008). Importantly, analogously to the CD_5 index, the CD_{10} index of papers and patents with zero references is either exactly equal to one (if they receive at least one citation within ten years after publication), or remains undefined. **a**, The distribution of the CD_{10} index for *SciSciNet*, created using the binwidth parameter in *seaborn 0.11.2*. Here again, the largest data points are hidden. **b,** The correct histogram of the underlying dataset. A peak at $CD_{10} = 1$ is revealed, corresponding to $9,187,034$ additional papers. **c,** The time evolution of the average CD_{10} index. When dropping the outliers with $CD_{10} = 1$, the decline in disruptiveness is negated. We find that 98% of the $CD_{10} = 1$ papers make zero references, consequently their exclusion impacts the data similarly. Therefore, our claim that papers with a CD index equal to one and zero references are driving the decline in the disruptiveness of scientific knowledge over time is unlikely to be dependent on the size of the forward citation window, which is used to calculate the respective CD index. The shaded bands correspond to 95 % confidence intervals. **d–f,** The equivalent plots for *PatentsView* revealing $153,027$ patents with $CD_{10} = 1$. When dropping the outliers with $CD_{10} = 1$, the decline in disruptiveness reduces substantially. We find that 87 % of the $CD_{10} = 1$ patents make zero references, consequently their exclusion impacts the decline similarly.

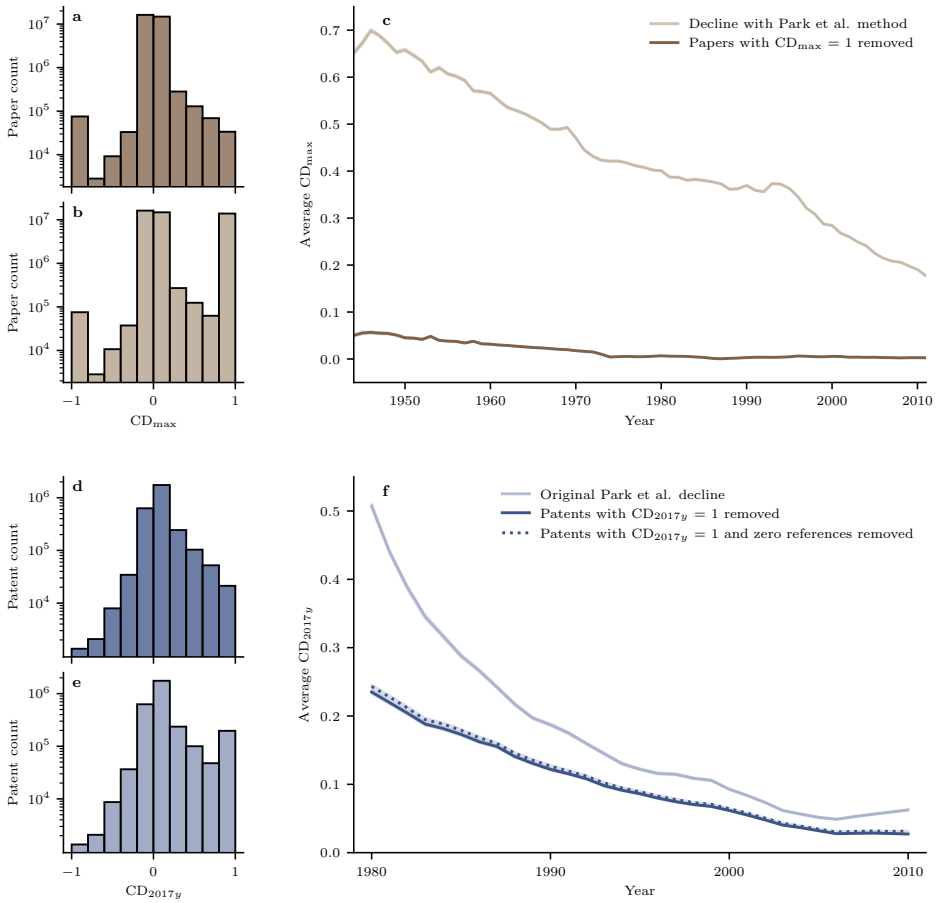

Fig. S7 | Distribution of the $\mathrm{CD_{max}}$ index with vs without the hidden outliers and its impact on the disruptiveness for the *SciSciNet* and the *PatentsView* data source. This figure displays the distribution and time average of the $\mathrm{CD_{max}}$ index (computed over a maximum forward citation window) for the *SciSciNet* [4] data source (with $45,564,829$ papers) and the *PatentsView* data source (with $3,011,723$ patents). In the notation of Park et al. [1], we have $\mathrm{CD_{max}} = \mathrm{CD}_{2022y}$ for papers and $\mathrm{CD_{max}} = \mathrm{CD}_{2017y}$ for patents. Importantly, the $\mathrm{CD_{max}}$ index of papers and patents with zero references is still either exactly equal to one (if they receive at least one citation), or remains undefined. **a**, The distribution of the $\mathrm{CD_{max}}$ index for *SciSciNet*, created using the binwidth parameter in *seaborn 0.11.2*. Here again, the largest data points are hidden. **b,** The correct histogram of the underlying dataset. A peak at $\mathrm{CD_{max}} = 1$ is revealed, corresponding to $13,864,845$ additional papers. **c,** The time evolution of the average $\mathrm{CD_{max}}$ index. When dropping the outliers with $\mathrm{CD_{max}} = 1$, the decline in disruptiveness is negated. We find that all of the $\mathrm{CD_{max}} = 1$ papers make zero references, consequently their exclusion is not shown separately. The shaded bands correspond to 95% confidence intervals. **d–f,** The equivalent plots for *PatentsView* revealing $175,190$ patents with $\mathrm{CD}_{2017y} = 1$. When dropping the outliers with $\mathrm{CD}_{2017y} = 1$, the decline in disruptiveness reduces substantially. We find that 94 % of the $\mathrm{CD}_{2017y} = 1$ patents make zero references, consequently their exclusion impacts the decline similarly.

S6. Normalized CD_5 indices

In this section, we analyse the distribution of the normalized CD_5 indices and the impact of the hidden outliers on the perceived temporal decline. We do this for four data sources: *Web of Science*, *PatentsView*, *SciSciNet* [4] and the *DBLP-Citation-network V14* [9]. The two normalized CD index variants are the ones used by Park et al. [1] (Extended Data Fig. 8**a,d** in [1]). Both variants adjust the term N_R (resp. N_k in the notation of Park et al. [1]) in the definition of the CD index, i.e. they modify the part of the definition that refers to follow-up work that does not cite the focal paper or patent itself but at least one of its references (Supplementary Equation S1 and S2).

- Paper (resp. patent) normalized CD index: replace N_R with $\max(N_R - \#\text{ref}, 0)$.
- Field x year normalized CD index: replace N_R with $\max(N_R - \overline{\#\text{ref}}(\text{field, year}), 0)$.

Here, #ref denotes the number of references of the focal paper (resp. patent), $\overline{\#\text{ref}}$(field, year) denotes the average number of references per year and field, and $\max(\cdot, \cdot)$ denotes the maximum between two values. We show our analysis in Fig. S8.

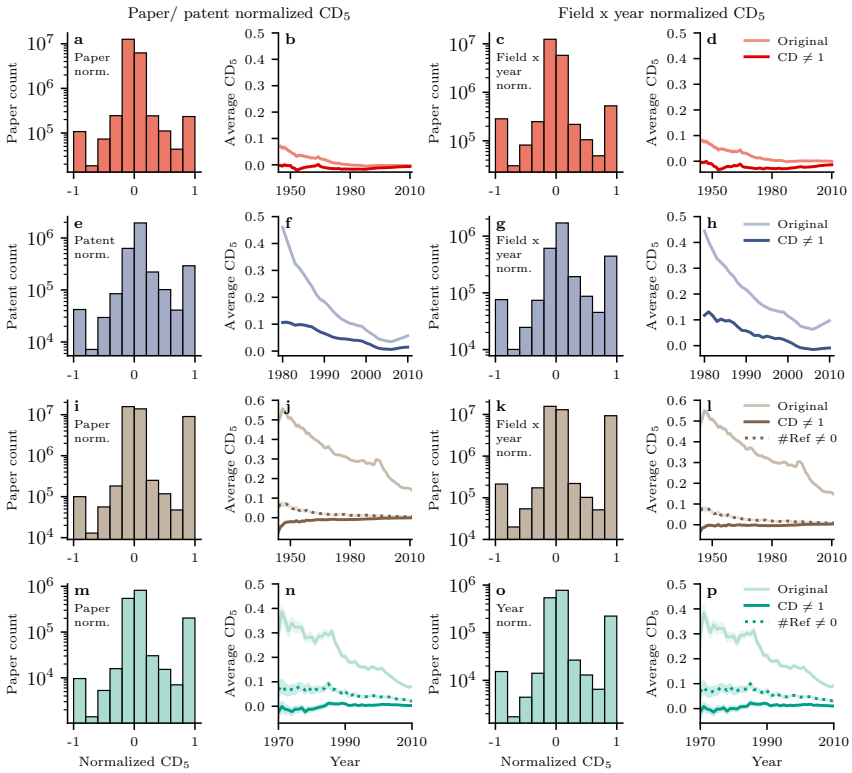

Fig. S8 | Distribution of normalized CD_5 indices and the impact of the hidden outliers on the perceived temporal decline. This figure shows how the hidden outliers are driving the decline for two normalized CD index variants used by Park et al. [1] (Extended Data Fig. 8**a,d** in [1]) across four different data sources: *Web of Science*: 19,927,359 (resp. 19,743,919) papers (**a**, **b**, **c**, **d**), *PatentsView*: 3,396,624 (resp. 3,270,187) patents, (**e**, **f**, **g**, **h**), *SciSciNet*: 39,473,940 (resp. 38,793,453) papers, (**i**, **j**, **k**, **l**) and the *DBLP-Citation-network V14*: 1,651,398 (resp. 1,623,775) papers, (**m**, **n**, **o**, **p**). Counts differ between the two normalized CD indices since they can be undefined for different papers and patents. The normalized variants adjust the term N_R (resp. N_k in the notation of Park et al. [1]; see also Supplementary equation S2) in the definition of the CD index, i.e. they modify the part of the definition that refers to follow-up work that does not cite the focal paper or patent itself but at least one of its references. A detailed definition of both variants can be found in the introduction of the present section. It is worth noting that for both normalized variants, papers and patents that make zero references have a value that either is exactly equal to one (if they are cited at least once) or remains undefined. Panels **a**, **e**, **i**, **m** reveal a peak at one for the paper (resp. patent) normalized CD index across all aforementioned data sources. Panels **b**, **f**, **j**, **n** show the time evolution of the average paper (resp. patent) normalized CD_5 index for papers. When dropping the hidden outliers with normalized CD index exactly equal to one, the decline in disruptiveness completely disappears for the paper datasets and substantially reduces for the patent dataset. For the *SciSciNet* data source and the *DBLP-Citation-network V14*, we had access to sufficient metadata to also exclude papers that make zero references similarly impacting the decline. Shaded bands correspond to 95% confidence intervals. The remaining panels show the equivalent plots for the field x year normalized CD index. Being specific to the field of computer science, we show the year normalized CD index for *DBLP-Citation-network V14*.

S7. Random paper and patent samples

This section contains tables with additional information on the 100 randomly drawn papers and patents from the sample with zero references and at least one forward citation within five years after publication, which results in $CD_5 = 1$.

Table S2 Summary statistics of the randomly drawn papers and patents with $CD_5 = 1$ and zero references.

Papers	Ref. in PDF	Zero ref. in PDF	Conference	Technical report	List of abstracts	News		
	93%	7%	1%	1%	4%	1%		
Patents	Ref. in PDF	Zero ref. in PDF	US Pre-1976	US Post-1976	A1	E	Foreign	Other
	98%	2%	49%	1%	16%	1%	48%	39%

Note: This table contains the summary statistics of 100 randomly drawn papers and patents from the sample with zero references and at least one forward citation within five years after publication, which results in $CD_5 = 1$. For *SciSciNet*, we draw papers from the *pandas DataFrame* with zero references and at least one forward citation using a random seed equal to zero. Since we do not have access to all the PDF files, we have to draw 238 papers until we obtain 100 accessible PDFs. The random sample for patents is drawn from the *pandas DataFrame* provided by Park et al. [1] using a random seed equal to zero. We manually verified the PDF files (see Supplementary Tables S3, S4) for references and find that 93% of papers and 98% of patents make at least one reference. For papers, we have a more detailed view on the type of the seven documents that do not make any references to prior work. We find four papers which are part of a list of abstracts. Further, there is a conference paper which mentions prior work but does not make explicit references, one Nature news article, and one technical report in the transportation research record. For patents, we have a more detailed view on which types of references are missing and report the percentage of patents which contain at least one reference to a specific category. Note that each patent may contain references to multiple types of references. We find 49% of patents contain references that are missing due to truncation caused by pre-processing the data (pre-1976) [11], 16% due to not accounting for patent applications since the passage of the Inventor Protection Act of 1999 (A1) [11], and 48% and 39% due to not counting foreign patents and other publications, respectively. Only one patent (id 6552498) contains a reference to a US patent, which seems to be a bibliometric error, and one patent contains a reference to a reissue patent (E). Finally, we find that these papers and patents have a median of 17.5 and 5 references which corresponds to the median of 9 and 5 references for the full sample, respectively. We compute the median for papers only for PDF files where we are able to count the exact number of references, i.e. by excluding the "1+" (see Supplementary Table S3 for details on the occurrences of "1+" in the case of only partial accessibility of the PDF file or references in footnotes).

Table S3: Paper details.

Paper ID	Year	Link	Open access	#References in PDF	PDF available	Foot-notes	Conference	Technical report	List of abstracts	News
27985486	2008	Link	no	27						
53937866	1993	Link	yes	6						
98255911	1998	Link	yes	13						
150390299	1994	Link	yes	1+	partial	no				
180437480	2000	Link	yes	15						
189909736	2006	Link	yes	18						
288606890	1963	Link	yes	6						
574253104	2006	Link	yes	1+	partial	yes				
575352733	2011	Link	yes	1+	partial	no				
585396547	2000	Link	yes	1+	partial	no				
585713240	1991	Link	yes	0			no	yes	no	no
639863130	2007	Link	yes	11						
952410954	2011	Link	yes	68						
1411236576	1972	Link	yes	9						
1493574560	1958	Link	yes	18						
1495110460	2003	Link	yes	17						
1522126398	1977	Link	no	12						
1536891849	1976	Link	yes	0			yes	no	no	no
1557088112	2002	Link	yes	1+	partial	yes				
1575011135	2009	Link	yes	1+	full	no				
1633088188	1991	Link	yes	1+	partial	no				
1657310252	1987	Link	yes	0			no	no	no	yes
1813921254	1975	Link	yes	1+	partial	yes				

Note: This table continues on the next three pages and contains the details of 100 randomly drawn papers from the sample with zero references and at least one forward citation, which results in $CD_5 = 1$. The random sample is drawn from the *pandas DataFrame* with zero references and at least one forward citation using a random seed equal to zero. Since we do not have access to all the PDF files, we have to draw 238 papers until we obtain 100 accessible PDFs.

Paper ID	Year	Link	Open access	#References in PDF	PDF available	Foot-notes	Conference	Technical report	List of abstracts	News
1818446669	1990	Link	yes	25						
1856184614	1969	Link	yes	40						
1968249834	1995	Link	yes	53						
1968371615	1990	Link	yes	14						
1970177916	1993	Link	no	30						
1971682916	1996	Link	yes	1+	partial	no				
1975852712	1988	Link	no	25						
1977359555	2006	Link	yes	31						
1978870643	2011	Link	yes	28						
1982053324	1995	Link	yes	1+	partial	yes				
1985373775	1996	Link	yes	5						
1989942175	1990	Link	yes	6						
1991296186	1993	Link	yes	39						
1996355016	1999	Link	yes	1+	full	yes				
2007371250	2004	Link	no	0			no	no	yes	no
2008609890	1991	Link	no	0			no	no	yes	no
2010293280	1990	Link	yes	42						
2023511737	1999	Link	no	15						
2024560885	1975	Link	yes	3						
2027584436	2006	Link	no	13						
2040840436	2006	Link	yes	54						
2042086152	2010	Link	yes	1						
2050466013	1999	Link	no	31						
2057336774	1993	Link	yes	1+	partial	yes				
2060753453	1999	Link	yes	18						
2063676441	2006	Link	yes	6						
2069640366	1992	Link	yes	22						

Paper ID	Year	Link	Open access	#References in PDF	PDF available	Foot-notes	Conference	Technical report	List of abstracts	News
2081547537	1969	Link	yes	1+	partial	no				
2083436642	2010	Link	yes	26						
2083638940	1991	Link	yes	1+	partial	yes				
2086255442	2008	Link	yes	1						
2090072051	1991	Link	no	19						
2117751013	2004	Link	yes	35						
2123154878	2008	Link	no	0			no	no	yes	no
2123761965	1968	Link	yes	1+	partial	yes				
2139136071	2011	Link	yes	1+	full	yes				
2313058271	1973	Link	no	41						
2317934052	1949	Link	yes	1+	partial	yes				
2328439434	1969	Link	no	14						
2347015509	2009	Link	no	33						
2389311988	2010	Link	yes	32						
2405115874	1986	Link	yes	24						
2410752987	1989	Link	yes	182						
2412560019	1999	Link	yes	15						
2412811714	2000	Link	yes	32						
2418732624	1991	Link	yes	6						
2418744057	1987	Link	yes	61						
2419118622	2002	Link	yes	5						
2461652553	1997	Link	no	51						
2484809223	2006	Link	yes	1+	partial	yes				
2490486294	1974	Link	yes	36						
2503700092	1999	Link	yes	147						
2507383598	1988	Link	yes	1+	partial	no				
2748776472	2006	Link	yes	40						

Paper ID	Year	Link	Open access	#References in PDF	PDF available	Foot-notes	Conference	Technical report	List of abstracts	News
2886101548	2006	Link	yes	5						
3103741798	2002	Link	yes	37						
3168256524	1990	Link	yes	0			no	no	yes	no
3188737630	2009	Link	yes	1+	partial	no				
75971907	1986	Link	yes	2						
128280131	1990	Link	yes	1+	partial	no				
235473009	2010	Link	yes	1+	full	yes				
257670102	2009	Link	yes	17						
288441233	1994	Link	yes	3						
649906705	1978	Link	yes	1+	partial	no				
1004243738	1981	Link	yes	3						
1488058573	1996	Link	yes	23						
1497336836	1985	Link	yes	1+	full	yes				
1553022496	2007	Link	yes	1+	full	yes				
1816860758	2010	Link	no	28						
1957779105	2003	Link	yes	1+	partial	no				
1976019262	1997	Link	no	14						
1978957683	1990	Link	yes	46						
1989925682	2002	Link	no	11						
2016946988	1967	Link	yes	17						
2019380115	1998	Link	no	22						
2025260115	1998	Link	yes	16						
2039319856	1997	Link	no	37						

Table S4: Patent details.

Patent ID	Grant Year	Link	#References in PDF	US Pre-1976	US Post-1976	A1	E	Foreign	Other
4295044	1981	Link	5	yes	no	no	no	no	no
5022291	1991	Link	4	yes	no	no	no	yes	no
4827022	1989	Link	8	no	no	no	no	yes	yes
7720724	2010	Link	11	no	no	yes	no	yes	no
4236087	1980	Link	4	yes	no	no	no	no	no
7048812	2006	Link	9	yes	no	no	no	yes	yes
4800247	1989	Link	0	no	no	no	no	no	no
7067678	2006	Link	7	no	no	yes	no	no	yes
7684359	2010	Link	5	no	no	yes	no	yes	no
4230169	1980	Link	6	yes	no	no	no	no	no
5596016	1997	Link	9	no	no	no	no	no	yes
4331055	1982	Link	7	yes	no	no	no	yes	no
5140549	1992	Link	1	yes	no	no	no	no	no
4857516	1989	Link	7	yes	no	no	no	no	yes
7304789	2007	Link	8	yes	no	yes	no	yes	no
5175384	1992	Link	8	no	no	no	no	no	yes
5360716	1994	Link	15	no	no	no	no	yes	yes
6180695	2001	Link	2	no	no	no	no	yes	no
7294680	2007	Link	10	no	no	no	no	yes	yes
4359443	1982	Link	10	yes	no	no	no	no	no
5065744	1991	Link	3	no	no	no	no	yes	no
5568322	1996	Link	6	yes	no	no	no	yes	no
4492663	1985	Link	9	yes	no	no	no	no	no

Note: This table continues on the next three pages and contains the details of 100 randomly drawn patents from the sample with zero references and at least one forward citation, which results in $CD_5 = 1$. The random sample is drawn from the *pandas DataFrame* provided by Park et al. [1] with a random seed equal to zero.

Patent ID	Grant Year	Link	References in PDF	US Pre 1976	US Post 1976	A1	E	Foreign	Other
4485566	1984	Link	3	yes	no	no	no	yes	no
7592005	2009	Link	38	no	no	yes	no	yes	yes
5122204	1992	Link	2	no	no	no	no	yes	no
4234775	1980	Link	7	yes	no	no	no	no	no
6860666	2005	Link	3	yes	no	no	no	no	no
7551067	2009	Link	1	no	no	no	no	yes	no
7604351	2009	Link	9	no	no	yes	no	yes	yes
5342757	1994	Link	6	no	no	no	no	no	yes
4865771	1989	Link	1	no	no	no	no	yes	no
6967241	2005	Link	9	no	no	yes	no	yes	yes
4435552	1984	Link	6	yes	no	no	no	yes	yes
4242334	1980	Link	6	yes	no	no	no	no	yes
4409862	1983	Link	6	yes	no	no	no	yes	no
7762264	2010	Link	7	no	no	no	no	no	yes
5865057	1999	Link	5	no	no	no	no	yes	no
4664811	1987	Link	5	yes	no	no	no	no	no
5707145	1998	Link	14	yes	no	no	no	yes	no
7041814	2006	Link	19	no	no	no	no	no	yes
4583720	1986	Link	6	yes	no	no	no	yes	no
5581312	1996	Link	1	no	no	no	no	yes	no
7502698	2009	Link	17	no	no	no	no	yes	no
7659264	2010	Link	1	no	no	no	no	yes	no
7805331	2010	Link	6	no	no	yes	no	no	yes
4368763	1983	Link	5	yes	no	no	no	yes	no
5709729	1998	Link	6	no	no	no	no	yes	yes
6941475	2005	Link	2	no	no	yes	no	no	no

Patent ID	Grant Year	Link	References in PDF	US Pre 1976	US Post 1976	A1	E	Foreign	Other
4194052	1980	Link	1	yes	no	no	no	no	no
5020086	1991	Link	0	no	no	no	no	no	no
6027363	2000	Link	2	yes	no	no	no	yes	no
6544700	2003	Link	1	no	no	no	no	yes	no
5523292	1996	Link	30	no	no	no	no	yes	yes
4573098	1986	Link	4	yes	no	no	no	no	yes
4257538	1981	Link	5	yes	no	no	no	no	no
4522521	1985	Link	13	yes	no	no	no	no	yes
4285934	1981	Link	2	no	no	no	no	yes	yes
4441704	1984	Link	2	yes	no	no	no	no	no
5869693	1999	Link	4	no	no	no	no	no	yes
5294907	1994	Link	1	yes	no	no	no	no	no
6028059	2000	Link	12	no	no	no	no	yes	yes
4353031	1982	Link	4	yes	no	no	no	no	no
7633830	2009	Link	1	no	no	yes	no	no	no
4238282	1980	Link	3	yes	no	no	no	no	no
6403089	2002	Link	1	no	no	no	no	no	yes
4383530	1983	Link	7	yes	no	no	no	yes	no
4312553	1982	Link	3	yes	no	no	no	yes	no
6948043	2005	Link	3	no	no	yes	no	no	yes
4186183	1980	Link	7	no	no	no	no	no	yes
6583828	2003	Link	1	no	no	no	yes	no	no
7587403	2009	Link	10	no	no	yes	no	yes	yes
7741196	2010	Link	12	no	no	yes	no	no	no
4332530	1982	Link	4	yes	no	no	no	no	no
4532975	1985	Link	5	yes	no	no	no	yes	yes

Patent ID	Grant Year	Link	References in PDF	US Pre 1976	US Post 1976	A1	E	Foreign	Other
6290363	2001	Link	1	yes	no	no	no	no	no
4182612	1980	Link	10	yes	no	no	no	no	yes
4331405	1982	Link	3	no	no	no	no	yes	no
4269657	1981	Link	7	yes	no	no	no	yes	yes
7843049	2010	Link	4	no	no	no	no	yes	no
5597943	1997	Link	4	yes	no	no	no	no	yes
7848211	2010	Link	4	no	no	yes	no	yes	no
4656294	1987	Link	2	yes	no	no	no	no	no
4587999	1986	Link	16	yes	no	no	no	yes	no
5112987	1992	Link	3	no	no	no	no	no	yes
5514068	1996	Link	4	yes	no	no	no	no	no
7466424	2008	Link	8	no	no	yes	no	no	yes
7129097	2006	Link	7	no	no	yes	no	yes	yes
7030189	2006	Link	5	no	no	no	no	yes	yes
5364994	1994	Link	2	no	no	no	no	no	yes
4940736	1990	Link	2	yes	no	no	no	no	no
5438051	1995	Link	4	no	no	no	no	no	yes
6552498	2003	Link	1	no	yes	no	no	no	no
4628776	1986	Link	6	yes	no	no	no	yes	no
5659116	1997	Link	4	no	no	no	no	no	yes
6641266	2003	Link	5	yes	no	no	no	yes	no
4283043	1981	Link	14	yes	no	no	no	no	no
4336212	1982	Link	11	yes	no	no	no	yes	no
4438473	1984	Link	4	yes	no	no	no	no	no
4364053	1982	Link	4	yes	no	no	no	no	no

References

[1] Park, M., Leahey, E. & Funk, R. J. Papers and patents are becoming less disruptive over time. *Nature* **613**, 138–144 (2023).

[2] Funk, R. J. & Owen-Smith, J. A dynamic network measure of technological change. *Management science* **63**, 791–817 (2017).

[3] Waskom, M. Treat binwidth as approximate to avoid dropping outermost datapoints. (2023). https://github.com/mwaskom/seaborn/pull/3489.

[4] Lin, Z., Yin, Y., Liu, L. & Wang, D. Sciscinet: A large-scale open data lake for the science of science research. *Scientific Data* **10**, 315 (2023).

[5] Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B. Atypical combinations and scientific impact. *Science* **342**, 468–472 (2013).

[6] Hofmann, H., Wickham, H. & Kafadar, K. value plots: Boxplots for large data. *Journal of Computational and Graphical Statistics* **26**, 469–477 (2017).

[7] Wu, L., Wang, D. & Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature* **566**, 378–382 (2019).

[8] Lin, Y., Frey, C. B. & Wu, L. Remote collaboration fuses fewer breakthrough ideas. *Nature* **623**, 987–991 (2023).

[9] Tang, J. *et al. Arnetminer: Extraction and mining of academic social networks*, KDD '08, 990–998 (Association for Computing Machinery, New York, NY, USA, 2008).

[10] Ruan, X., Lyu, D., Gong, K., Cheng, Y. & Li, J. Rethinking the disruption index as a measure of scientific and technological advances. *Technological Forecasting and Social Change* **172**, 121071 (2021).

[11] Macher, J. T., Rutzer, C. & Weder, R. The illusive slump of disruptive patents. *arXiv preprint arXiv:2306.10774* (2023).