Title

The disruption index suffers from citation inflation: Re-analysis of temporal CD trend and relationship with team size reveal discrepancies

Journal

Journal of Informetrics, 19(1)

ISSN

1751-1577

Authors

Petersen, Alexander Michael
Arroyave, Felber J
Pammolli, Fabio

Publication Date

2025-02-01

DOI

10.1016/j.joi.2024.101605

Journal of Informetrics 19 (2025) 101605

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Research Paper

The disruption index suffers from citation inflation: Re-analysis of temporal CD trend and relationship with team size reveal discrepancies

Alexander Michael Petersen [a,*], Felber J. Arroyave [a], Fabio Pammolli [b]

[a] Department of Management of Complex Systems, Ernest and Julio Gallo Management Program, School of Engineering, University of California, Merced, CA 95343, USA

[b] Politecnico Milano, Department of Management, Economics and Industrial Engineering, Via Lambruschini, 4/B, 20156, Milan, Italy

ARTICLE INFO

Dataset link: 10.6071/M3G674

Keywords:
Disruption index
Innovation
Secular growth
Measurement bias
Quasi-experiment
Reproducibility analysis

ABSTRACT

Measuring the rate of innovation in academia and industry is fundamental to monitoring the efficiency and competitiveness of the knowledge economy. To this end, a disruption index (CD) was recently developed and applied to publication and patent citation networks (Wu et al., 2019; Park et al., 2023). Here we show that CD systematically decreases over time due to secular growth in research production, following two distinct mechanisms unrelated to innovation – one behavioral and the other structural. Whereas the behavioral explanation reflects shifts associated with techno-social factors (e.g. self-citation practices), the structural explanation follows from 'citation inflation' (CI), an inextricable feature of real citation networks attributable to increasing reference list lengths, which causes CD to systematically decrease. We demonstrate this causal link by way of mathematical deduction, computational simulation, multi-variate regression, and quasi-experimental comparison of the disruptiveness of PNAS versus PNAS Plus articles, which differ primarily in their lengths. Accordingly, we analyze CD data available in the SciSciNet database and find that disruptiveness incrementally increased from 2005-2015, and that the negative relationship between disruption and team-size is remarkably small in overall magnitude effect size, and shifts from negative to positive for team size ≥ 8 coauthors.

1. Introduction

Disruptive innovation refers to intellectual and industrial breakthroughs that sidestep conventional theory or practice by appealing to new value networks, to the extent that the disruptive entrants can quickly and unexpectedly overcome the competitive advantages characteristic of established incumbents (Christensen et al., 2015). In the case of scientific advancement, the process of disruptive innovation manifests as intellectual contributions that appeal to novel configurations of concepts and methods belonging to the knowledge network (Pan et al., 2018; Petersen, 2022; Yang et al., 2023), thereby substituting prior combinatorial knowledge à la Schumpeter's theory of creative destruction (Schumpeter, 1942). Against this backdrop, Funk and Owen-Smith (2017) recently developed an index for quantifying citation disruption (denoted by CD) according to the implicit value of intellectual attribution that is encoded within the local structure of citation networks, with the objective of identifying intellectual contributions that appeal to new streams of intellectual attribution while subverting established ones.

* Corresponding author.
E-mail address: apetersen3@ucmerced.edu (A.M. Petersen).

https://doi.org/10.1016/j.joi.2024.101605
Received 14 April 2024; Received in revised form 29 July 2024; Accepted 7 November 2024

Available online 16 November 2024
1751-1577/

Identifying the micro-level processes underlying disruption and quantifying its overall rate are fundamental to understanding scientific progress, and can in principle guide the management of institutions and policies that accelerate innovation. As such, the CD index has received considerable attention and inspired a significant volume of follow-up research. However, there is a growing literature that challenges the definition and application of CD to real scientific and patent citation networks (Bentley et al., 2023; Bornmann et al., 2020; Holst et al., 2024; Leibel & Bornmann, 2024; Macher et al., 2024; Petersen et al., 2024; Ruan et al., 2021). One stream of critique calls into question the long-term temporal decline in CD reported by Park et al. (2023), and the morose interpretation of its implications on the status and outlook of the scientific endeavor (Holst et al., 2024; Kozlov, 2023). In particular, Macher et al. (2024) identify a substantial number of missing patent citations in the data analyzed by Park et al. (2023), both at the beginning of their data sample and towards the end, which effectively reduces the number of backwards citations (i.e. references) and forward citations that were analyzed. Once correcting for the data omission, the re-analysis reveals an *increasing rate* of disruption in several patent domains central to the techno-informatic revolution of the last 30 years. This increase is consistent with the persistent increase in combinatorial innovation observed in science over the same period (Yang et al., 2023), and the general purpose technologies of the AI revolution (Eloundou et al., 2024) that have already disrupted the process of knowledge production across the sciences (Abramson et al., 2024; Davies et al., 2021; Krenn & Zeilinger, 2020; Merchant et al., 2023; Tshitoyan et al., 2019; Wang et al., 2023a).

Similarly, a recent independent re-analysis by Holst et al. (2024) shows that missing citations in the scientific publication data, which are more prominent in the early years of the sample, give rise to a substantial subsample of publications with 0 references – which by definition correspond to maximum disruption value of $CD = 1$. They show that the prevalence of these temporally biased data anomalies are entirely sufficient to generate the negative trend in CD reported by (Park et al., 2023); upon correcting for these anomalies, they show that CD trends insubstantial for both patents and publications. A third independent study on scientific publications by Bentley et al. (2023) also reports an accelerating rate of disruption at the end of their data sample after developing a weighted variant of CD that accounts for temporal shifts in the connectivity of real citation networks. In addition to these studies reporting an increasing rate of innovation according to temporal patterns in the citation network, a complementary approach based upon measuring combinatorial innovation in the knowledge network also reports a persistent innovation rate over time (Petersen, 2022).

A second stream of critique focuses on how CD is defined, and the implications of its mathematical formulation on its reliability in bibliometric analysis (Bornmann et al., 2020; Leydesdorff et al., 2021; Petersen et al., 2024; Ruan et al., 2021; Wu & Wu, 2019). Taken together, these considerations further call into question a number of studies reporting statistical relationships between CD and various covariates related to research production and team assembly – such as career productivity, team size, citation impact, and the geographic dispersion of team members (Li et al., 2024; Lin et al., 2023a; Park et al., 2023; Wang et al., 2023b; Wu et al., 2019) – most notably because each of these covariates also grows with time. Statistical relationships between CD and other time-dependent covariates are susceptible to omitted variable bias, which is a formidable source of measurement error in statistical analysis. As such, the connection between these two streams is the role of secular growth, which manifests as a temporal bias that underlies the data artifacts generating declining trends in CD and the susceptibility of CD to measurement error due to its non-linear dependence on the structure and rate of backwards citations.

Against this backdrop, here we contribute to these two streams by demonstrating how systematic measurement bias deriving from the inextricable densification of empirical citation networks, combined with omitted variables capturing confounding shifts in scholarly citation practice, contributes to the mis-measurement of a "decline in disruptiveness" (Park et al., 2023). Our critique is centered upon the role of reference list in the definition of CD, and the implications of multifold increases in reference list lengths over the last half century, which is a fundamental source of 'citation inflation' (CI) (Petersen et al., 2018). Specifically, we apply four complementary methodologies that expose the underlying bias in CD definition and its application – deductive quantitative reasoning, computational modeling, a quasi-experimental test, and multivariate regression – the results of which are consistent with a companion study (Petersen et al., 2024) based upon different data sources and regression model specification.

In order to test and validate the CI hypothesis – that increasing reference list lengths confound the measurement of trends in CD and covariate relationships – we developed a quasi-experiment based upon the entire corpus of research published in *Proceedings of the National Academy of Sciences of the United States of America* (PNAS) over the 5-year period 2011-2015. Our identification strategy is based around comparing research articles published in the traditional format (print and online publication) to those published as long-form *PNAS Plus* articles (online-only publication) (Schekman, 2010; Verma, 2012) – as these two publication formats are nearly indistinguishable, aside from the longer reference list lengths of *PNAS Plus* articles. We conclude with a large-scale multivariable regression analysis of 7.8 million articles from 1995-2015, which accounts for the data quality and measurement biases identified, thereby improving upon the methodological designs employed in Park et al. (2023); Wu et al. (2019). Results show that (a) the net effect size of temporal and team-size trends are at the level of noise and therefore inconsequential to science and innovation policy; and (b) by including appropriate controls and focusing on the metric itself instead of percentile values, the relationship between CD and team size is instead increasing.

2. Background & related literature

2.1. Definition of CD and its susceptibility to secular growth

The disruption index CD_p (Funk & Owen-Smith, 2017; Park et al., 2023) measures the degree to which an intellectual contribution p (e.g. a patent or academic research publication) supersedes the sources cited in its reference list, denoted by the set $\{r\}_p$. The

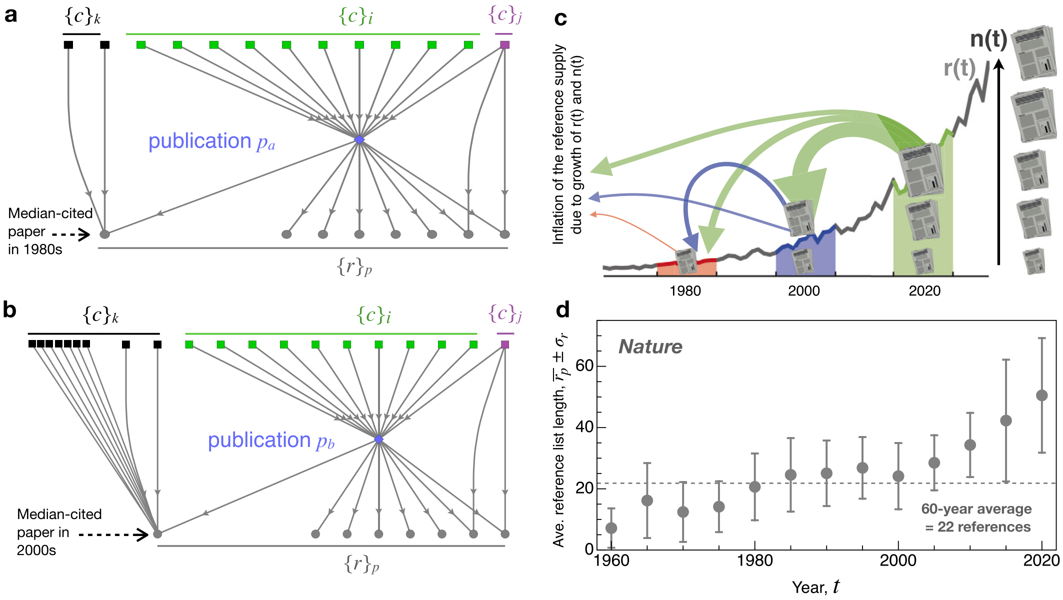

Fig. 1. Citation inflation is an inextricable feature of citation networks. The disruption index CD_p is calculated according to three non-overlapping subsets of $\{c\}_p = \{c\}_i \cup \{c\}_j \cup \{c\}_k$, of sizes N_i, N_j and N_k, respectively. **(a,b)** Schematic of the citation network sub-graph contributing to the calculation of the disruption index for two papers that differ only in the connectivity of the single reference contributing to N_k. Moreover, in order to convey the magnitude and impact of secular growth as it manifests on real citation networks, the subset $\{c\}_k$ for publication p_a is characteristic of citation rates in the 1980s, whereas for p_b it is characteristic of the 2000s. Consequently, $CD_a = 0.69$ and $CD_b = 0.45$, corresponding to a 35% decrease in CD attributable to 20 years of increasing citation network density. **(c)** Schematic illustrating the inflation of the reference supply: both the annual publication rate $n(t)$ (comprised of increasingly variable article lengths) and the number of references per publication, $r(t)$, have grown exponentially over time. Consequently, the observed densification is both within- and across-generation, such that older publications can receive more citations from present day research than from contemporaneous research due to secular growth. **(d)** Citation inflation even affects journal with relatively small change in $n(t)$, such as traditional print journals like *Nature*, which have witnessed 7-fold increases in reference list lengths over the last 60 years.

argument for CD_p is that if future contributions cite p but do not cite members of $\{r\}_p$, then p plays a disruptive role in the citation network. As such, disruption can be inferred according to the local structure of the subnetwork $\{r\}_p \cup p \cup \{c\}_p$ that includes the set of citing nodes $\{c\}_p$ connecting to either the focal node p or any member of $\{r\}_p$. According to its definition (Funk & Owen-Smith, 2017; Park et al., 2023) reformulated as a ratio (Wu et al., 2019), CD_p is calculated by identifying three non-overlapping subsets of $\{c\}_p = \{c\}_i \cup \{c\}_j \cup \{c\}_k$, of sizes N_i, N_j and N_k, respectively – see **Fig. 1**(a,b) for a schematic illustration. In practice, a citation window (CW) is used to temper the effects of right-censoring bias, such that only citations occurring within a CW-year period are included in the subnetwork $\{c\}_p$. In what follows, we employ a CW = 5-year window denoted by $CD_{p,5}$, as in prior research (Park et al., 2023; Wu et al., 2019); however, the fundamental issues with the definition of CD are independent of CW (Petersen et al., 2024), and so for brevity we represent the general definition by CD_p.

The subset i refers to members of $\{c\}_p$ that cite the focal p but do not cite any elements of $\{r\}_p$, and thus measures the degree to which p disrupts the flow of attribution to members of $\{r\}_p$. The subset j refers to members of $\{c\}_p$ that cite both p and $\{r\}_p$, measuring the degree of consolidation that manifests as triadic closure in the subnetwork (i.e., triangles formed between $\{r\}_p$, p, $\{c\}_j$). The subset k refers to members of $\{c\}_p$ that cite $\{r\}_p$ but do not cite p. As such, Wu et al. (2019) show that CD_p can be calculated as the ratio

$$CD_p = \frac{N_i - N_j}{N_i + N_j + N_k} = \frac{CD^{\text{nok}}}{1 + R_k}\,, \tag{1}$$

where the second equivalence is a simple re-organization of the equation to highlight the extensive quantity $R_k = N_k/(N_i + N_j) \in [0, \infty)$, which measures the rate of extraneous citation. Such re-organization facilitates deducing the scaling behavior of CD associated with the number of articles $n(t)$ and the average reference list length per article $r(t)$ per year t.

According to the network growth properties, we can approximate how N_i, N_j and N_k grow using basic scaling arguments based upon the two fundamental sources of growth, $n(t)$ and $r(t)$. First consider publications in the reference list of p_a in **Fig. 1**(a) that are unconstrained by the first-order citation network $\{c\}_p$ around the focal publication p from year t. Since most citations are made between papers with relatively small publication-year differences (Pan et al., 2018), then the dominant first-order contribution to N_k is the contemporaneous total citation rate, $C(t) = n(t)r(t)$. Now considering the set $\{c\}_i \cup \{c\}_j$, since these publications conditionally cite p and can only cite p once, then $N_i + N_j$ grows as $n(t)$. Hence, because $N_k \sim n(t)r(t)$ and $N_i + N_j \sim n(t)$ then $R_k = N_k/(N_i + N_j) \propto r(t)$; see (Petersen et al., 2024) for empirical validation and **Fig. S1** in the *Supplemental Materials* for computational validation.

Consequently, because the ratio $CD^{\text{nok}} = (N_i - N_j)/(N_i + N_j) \in [-1, 1]$ is an intensive measure, CD_p features a numerator that is bounded and a denominator that is unbounded. As such, $CD_p(t)$ converges over time to 0 because the denominator grows proportional

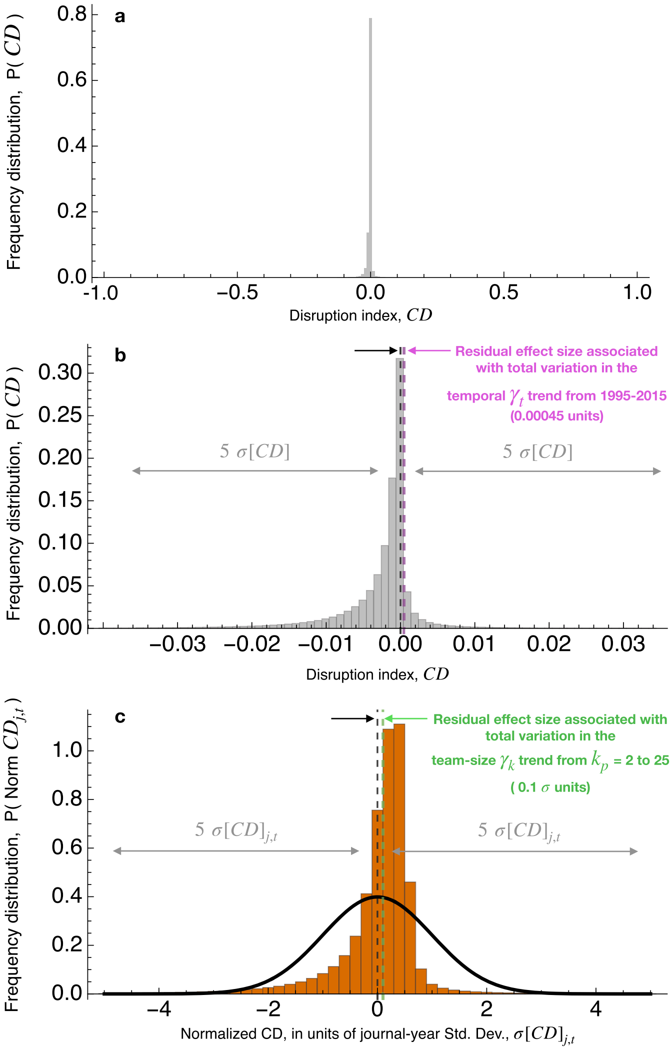

Fig. 2. Descriptive statistics: extremely concentrated and leptokurtic distribution of CD. Publication data were obtained from the *SciSciNet* open data repository (Lin et al., 2023b), selecting journal articles published in the 21-year range 1995-2015, and further selecting only the articles with reference list lengths in the range $10 \le r_p \le 200$. This assures that we are not basing results upon editorials, comments, and other articles that do not represent substantial research products, and avoids the susceptibility of CD to small data sample fluctuations, e.g. as shown by Holst et al. (2024) that $CD = N_i/N_i = 1$ if $r_p = 0$ and $c_p > 0$, which occurs surprisingly frequently with a bias towards early years. **(a)** The frequency distribution $P(CD_{p,5})$ is extremely leptokurtic. **(b)** Zooming in on the $\pm 5\sigma$ range of the $CD_{p,5}$ distribution shows that 95.5% of the data are within $\pm 2\sigma$ of the average CD value. The deviation between the vertical black and magenta lines shows the relatively small amount of variation in CD attributable to the temporal trends shown in **Fig. 5**(a), which shows γ_t from 1995-2015 (a total of 0.00045 units of CD, corresponding to 0.06σ in terms of the standard deviation of CD). **(c)** Frequency distribution $P(\text{Norm}CD_{p,5})$ calculated for the normalized metric Norm$CD_{p,5}$ defined in Eq. (2), and plotted over the $\pm 5\sigma$ range. The deviation between the vertical black and green lines shows the 0.09σ effect size attributable to the team-size trends shown in **Fig. 5**(b). For visual comparison, the solid black curve represents a $N(0,1)$ normal distribution, showing that CD is still leptokurtic after accounting for journal-year average and variation scale with Norm$CD_{p,5}$.

$r(t)$, which itself grows exponentially, $r(t) \propto \exp[g_r t]$ with $g_r = 0.018$ corresponding to a roughly 2% annual growth rate (Pan et al., 2018; Petersen et al., 2018). This convergence of $CD_p(t) \to 0$ as $r(t)$ increases over time is conveyed by the very narrow empirical frequency distribution $P(CD)$ calculated for a large sample of 7.66 million publications from the 21-year period 1995-2021 – see **Fig. 2**(a).

2.2. Citation inflation: a measurement bias deriving from secular growth

Citation inflation (CI) refers to the exponential growth of citations produced via the secular growth of the scientific endeavor (Petersen et al., 2018), and has a natural counterpart in patenting (Huang et al., 2020). **Fig. 1(c)** illustrates how CI arises through the combination of increasing reference lists, denoted by $r(t)$, and increasing publication (or patenting) rates, denoted by $n(t)$, which significantly increases the density of citation networks over time. In the case of scientific research production, empirical growth rates estimated from the entire Clarivate Analytics Web of Science citation network show that total volume of citations generated by the scientific literature, $C(t)$, grows exponentially with annual rate $g_C = g_n + g_r = 0.051$; hence, the number of links in the citation network is growing by roughly 5% annually, corresponding to a doubling period of just $\ln(2)/g_C = 13.6$ years (Pan et al., 2018).

Moreover, as references tend to increasingly extend further back in time, the impacts of CI are not constrained to contemporaneous layers of the citation network, but instead are cross-generational. As such, the increasing density of citation networks manifests at both the source (reference) and destination (citation) of each new link, which is a temporal bias that is challenging to neutralize in the development of standardized network metrics. For example, in the 1980s the median-cited paper received 2 citations within 5 years of publication; however by the 2000s, this nominal quantity increased to 9 (Pan et al., 2018). In addition, there has been a paradigm shift towards online publishing that has facilitated greater publication volumes, faster publication times, and longer articles with longer reference lists. Take for example the journal Nature for which $n(t)$ has been roughly constant over the last 60 years: in 1970 the average number of references per articles was $\overline{r_p} = 7$; by 2000 $\overline{r_p}$ increased to 24; and by 2020 $\overline{r_p} = 51$, a 7-fold increase over the 60-year period – see **Fig. 1**(d) and **Fig. S2**.

2.3. Issues with prior works analyzing trends in CD

Park et al. (2023) develop four robustness check approaches: comparison with alternative definitions of CD, normalization of N_k in CD, regression adjustment, and synthetic randomization of the citation network. Yet each is susceptible to either data quality issues that are temporally biased towards early years, or measurement bias and omitted variable bias associated with the definition of CD.

First, because alternative disruption variants, namely CD^{nok} and CD^{*} (Bornmann et al., 2020; Leydesdorff et al., 2021), are constructed around similar ratios, they are also susceptible to data quality issues as well as CI. However, it is notable that the alternative indices are less susceptible to CI, and indeed their trends shown in Extended Data Fig. 7 of Park et al. (2023) are markedly less prominent than what is shown for CD. In the particular case of CD^{nok}, the location of the average value is large enough that a vast majority of papers are classified as disruptive – which begs for metric validity improvement. Moreover, in our companion study focusing upon a computational model, we show that even the tempered decline in CD^{nok} can be attributed to CI (Petersen et al., 2024).

Second, in order to attenuate the effect of CI, Park et al. (2023) develop both 'paper' and 'field x year' normalized variants of CD by modifying the factor N_k in Eq. (1). In the first case, they replace N_k with $N_k - r_p$. However, according to scaling behavior arguments, since $N_k \sim n(t)r(t)$ then $(N_k - r_p) \sim (n(t)r(t) - r_p) \approx (n(t) - 1)r(t)$, which does not generate the intended consequence. Moreover, the reduction of N_k by r_p still renders these normalized variants susceptible to the scenario exhibited in **Fig. 1**(a,b), whereby citing just a single highly-cited paper causes N_k to vastly exceed the difference $N_i - N_j$ in the numerator of CD, such that $CD_p \to 0$. In the second approach, using the field-year normalized variant, the average $r(t)$ for papers from the same field and year is subtracted. This, however, results in the same issue, $N_k - r(t) \sim n(t)r(t) - r(t) = (n(t) - 1)r(t)$.

Third, the regression adjustment implemented by Park et al. (2023) is poorly documented, as there is no model specification; moreover, Extended Table 8 only shows the estimated coefficients for the year indicator variable, and does not show the estimates for other controls. And according to Extended Data Table 1 in Park et al. (2023), their model specification does not incorporate available publication-level factors that co-vary with CD_p, namely r_p, c_p and the number of coauthors, k_p.

And finally, robustness checks based upon rewired citation networks are insufficient since the degree-preserving randomization holds constant r_p or c_p. Hence, this randomization scheme can only be expected to attenuate biases attributable to correlated citation behavior – contrariwise, biases deriving from data quality issues and CI can be expected to persist. Moreover, because shuffling the citation network reduces the rate of triadic closure to random chance, then this null model essentially converts all N_j links into N_i links. Consequently, the expected randomized value is $CD_p^{\mathrm{Rand}} = ((N_i + N_j) - 0)/(N_i + N_j + N_k) = 1/(1 + R_k) > 0$, which is positive definite and converges to 0 as R_k increases. Park et al. then calculate a Z-score comparing the real and randomized values, $Z_p = (CD_p - CD_p^{\mathrm{Rand}})/\sigma[CD_p^{\mathrm{Rand}}]$, and plot the average value over time in Extended Data Fig. 8 (Park et al., 2023). Their results show extremely negative Z-score values (upwards of 2σ effect sizes). These deviations are also methodological artifacts: since $CD_p^{\mathrm{Rand}} > 0$, then all papers with $CD_p < 0$ deterministically yield $Z_p < 0$; moreover, the standard deviation $\sigma[CD_p^{\mathrm{Rand}}]$ is extremely small because the chances of the randomization producing triadic closure are extremely small. Hence, with little variation to work with around a systematically small value $CD_p^{\mathrm{Rand}} \approx 1/(1 + R_k) \sim 1/r(t)$, this randomization approach vastly underestimates the intrinsic scale of variation, i.e., $\sigma[CD_p^{\mathrm{Rand}}] \ll \sigma[CD_p]$.

In a different study on the relationship between CD and team size, Wu et al. (2019) also develop robustness checks based upon multi-variate regression. However there is no clear model specification provided in their Supplementary Table 4; hence, in addition to omitting c_p and r_p, it is unclear how they controlled for publication year. Moreover, the majority of their analysis is based upon descriptive trend analysis using *percentile* values of CD. This mapping of nominal CD values to percentiles obfuscates the extremely narrow distribution of CD – see **Fig. 2**(a). At the same time, this modification of dependent variable generates the appearance of considerable effect sizes. Because most publications are concentrated around relatively small CD values, a small idiosyncratic shift in CD will generate disproportionately large shifts in the percentile value.

A third study analyzes the relationship between CD and collaboration distance among coauthors (Lin et al., 2023a). This analysis also omits c_p and r_p from their multivariate robustness check (see Extended Data Table 1), and instead categorizes papers as being before or after 2000 (a crude temporal control) and being solo-author or not (a crude team size control). As an example of persistent negligence for confounding factors, Lin et al. write: "For example, the 1953 paper on DNA by Watson and Crick is among the most disruptive works (D = 0.96, top 1%), whereas the 2001 paper on the human genome by the International Human Genome Sequencing

Consortium is highly developing (D = -0.017, bottom 6%).” Yet these papers are from vastly different socio-technological eras, with the former produced by two coauthors citing $r_p = 6$ prior works, where the latter is attributed to 200+ coauthors and cites $r_p = 452$ prior works.

Additionally, results reported by Lin et al. (2023a) are based upon the relative rates of $CD > 0$ versus $CD < 0$. This dependent variable is simply the sign of CD, which only depends on the numerator difference, $N_i - N_j$. While this choice may at first appear to be less susceptible to CI, it is still susceptible to the data quality issues identified by (Holst et al., 2024; Macher et al., 2024), as well as confounding trends in the rate of triadic closure attributable to shifts in self-citation and other correlated citation behaviors (Tahamtan & Bornmann, 2018), which are more prominent in larger teams – and larger teams are more likely to be extended across larger distances.

In response to these issues, two streams of critique have emerged regarding research on CD – one regards data quality issues and the other regards methodological choices. Regarding the former, citation networks based upon publication and patent data are susceptible to missing references, citations, and the mis-classification of non-research oriented content (e.g. editorials) as the products of dedicated research. These data quality issues are more frequent for older publications, and less so for newer ones, as the modern publication industry benefits from information system features that were not available in the past (e.g. Digital Object Identifiers, standard typesetting tools such as LaTex, and web-based publication). As demonstrated by Holst et al. (2024), the frequency of publications with 0 references is highly concentrated during early years. They show that this systematic bias in data quality contributes significantly to the decline in CD, since papers with $r_p = 0$ correspond to maximum disruption, $CD_p = (N_i - 0)/(N_i + 0 + 0) = 1$. Similarly, Macher et al. (2024) show that left-censoring bias in US patent data means that patents from early years are artificially missing references to patents before the starting date of the dataset; upon correcting for these omitted references, which increases r_p closer to their true value, then the decline in CD for patents is greatly reduced.

A second stream of critique regards the methodological choices, e.g. omitted variables and the susceptibility of CD to secular growth. By way of example, Bentley et al. (2023) modify the definition of CD_p to account for CI according to both the number of publications $n(t)$ and the total number of citations produced, $C(t) = n(t)r(t)$. Their re-analysis reveals an increasing weighted CD from the early 1990s through 2013. Bentley et al. (2023) also critique the natural language analysis by Park et al. (2023), noting that inferences based upon word usages are also susceptible to secular trends in usage frequency and semantic shift (Hamilton et al., 2016; Kristiansen, 2008).

To summarize, several recent analyses report findings of the following form: as X increases, CD decreases (where X is time, individual publication rate, team size, nominal citations, and collaboration distance) (Li et al., 2024; Lin et al., 2023a; Park et al., 2023; Wang et al., 2023b; Wu et al., 2019). Accordingly, we conjecture that any variable $X(t)$ that increases over time will generate correlations of this pattern – yet the degree to which such correlations survive confounders and whether they represent significant effect sizes is a more intriguing matter. To this end, here we seek to consolidate a growing number of critiques – first by addressing methodological issues, and concluding with a regression framework that also addresses the data quality issues.

3. Methods

3.1. Computational simulation model – testing the CI hypothesis

Computational ‘toy models’ are designed to capture the essential parameters underlying observed variation, while neglecting those features that are perceived to be weakly related. However, an unavoidable limitation to such approaches is determining what exactly are the essential parameters. Motivated by these modeling principles, we exploit a parsimonious citation network growth model that does not account for various sources of heterogeneity underlying scientific publication, such as team size and reference list lengths. Instead, all synthetic publications in our model from the same year have the same number of references, $r_p = r(t) \equiv$ average reference list length. In this way, we can rule out the intra-annual variation in r_p as a source of the effect we are seeking to understand and isolate.

The computational model does account for latent features of scientific production, in particular the exponential growth of $n(t)$ and $r(t)$. As such, we can use our model to re-analyze the temporal trends in CD reported by Park et al. (2023), using two types of synthetic networks generated by our model – one generated with CI and one without CI – that are otherwise statistically identical by construction. Our citation network model belongs to the class of growth and redirection models (Barabasi, 2016; Krapivsky & Redner, 2005) and implements stochastic link dynamics that mimic preferential attachment (Barabasi, 2016). The statistical properties of the citation networks generated by the model were tested and validated against comprehensive real citation data from the Web of Science (Pan et al., 2018). This model reproduces a number of statistical regularities established for real citation networks – both structural (e.g. a log-normal citation distribution (Radicchi et al., 2008)) and dynamical (e.g., increasing reference age over time (Pan et al., 2018); exponential citation life-cycle decay (Parolo et al., 2015; Petersen et al., 2014)).

Network growth in this model is governed by two complementary citation mechanisms that can be controlled by tunable parameters: (i) direct citation and (ii) redirected citation (Pan et al., 2018; Petersen et al., 2024). The second mechanism (ii) controls the rate of triadic closure in the synthetic citation network, thereby capturing correlated shifts in scholarly citation practice, such as citation trails illuminated by web-based hyperlinks that make it easier to find and cite prior literature; and self-citation among individuals and journals aimed at increasing their prominence in the attention economy. As such, the latter mechanism models the ‘consolidation’ measured by N_j in CD_p, which explicitly measures the number of citations that feature triadic closure. In related work focusing on the details of this computation model (Petersen et al., 2024), we find that CI has a stronger role in explaining the decline in $CD_5(t)$ than the redirection mechanism, and so in what follows we mainly focus on the effects of CI.

Accordingly, here we focus on the network growth parameters that determine the rate of CI, which thereby facilitates measuring the impact of secular growth on two trends: (a) the decline in the average $CD_{p,5}$ value in year t, denoted by $CD_5(t)$; and (b) the CI scaling hypothesis $R_k(t) \propto r(t)$ connecting the rate of extraneous citations featured in the denominator of Eq. (1) to the average number of references per paper. Note that the increase in $r(t)$ manifests from secular growth as well as shifts in scholarly citation practice. For example, papers with more authors tend have longer reference lists (Petersen et al., 2024), partly because larger teams tend to write longer papers, but also because there are more authors seeking to benefit from self-citation.

We use this computational model to test the CI hypothesis by generating two distinct network ensembles, each comprised of 10 random networks grown over $t = 1 \dots T$ periods (representative of publication years), terminating the network growth for $T \equiv 150$. Each network realization is seeded with common initial conditions – e.g. the initial cohort features $n(1) = 30$ nodes, each with $r(1) = 5$ references. These synthetic citation networks are available for cross-validation, and can be used to develop alternative scientometrics that are less biased by CI (see Data Accessibility statement).

The first ensemble incorporates the empirical rates $g_n = 0.033$ and $g_r = 0.018$ for the entire $T \equiv 150$ periods, with networks reaching a final size of $N = 125,270$ nodes and 5,948,492 links with $r_p \equiv r(t = 150) = 73$ after 150 periods. The second ensemble features the same g_n and thus reaches the same final size of $N = 125,270$ nodes. However, the growth of $r(t)$ is suddenly quenched at $T^* = 108$ to $g_r = 0$, such that $r(t) = 34$ for $t \geq T^*$, which effectively 'turns off' CI attributable to growing reference lists – see **Fig. S1**(a,b). We note that Pan et al. (2018) use this computational model to instead explore the implications of a sudden increase in $r(t)$, i.e. simulating the emergence of online-only mega-journals and their impact on the citation network.

3.2. *Multi-variate empirical analysis of $|CD_{p,5}|$ – testing the CI hypothesis*

In order to further test the CI hypothesis in an empirical setting, we resort to the following simplified scenario. We analyze the absolute value $|CD_{p,5}|$ so that the regression model specification is not confounded by the two different mechanisms that can generate reductions in CD – one relating to the numerator of Eq. (1) which depends on the rate of triadic closure in the network, and the other relating to the denominator which directly results from CI. After confirming the CI hypothesis, we then analyze $CD_{p,5}$ in the following section in order to further test the relationship between disruption and covariates.

To empirically test the CI hypothesis, we exploit the 2011 launch of a strategic publishing model developed by the journal PNAS, consisting of a long-form online-only publishing option – called *PNAS Plus* – to complement its traditional print option (Schekman, 2010). Article submissions were not processed, reviewed or prioritized according to the author-designated print option (Verma, 2012), and so publications in these two formats are satisfactory counterfactuals for testing whether publications with larger r_p are biased towards smaller CD_p. To demonstrate this causal link, we juxtapose the two sets of articles that are otherwise indistinguishable, on average, aside from *PNAS Plus* articles having longer reference lists. See **Fig. S3** for comparison of the two subsamples across various characteristics.

To facilitate reproduction, we base our empirical analysis on pre-computed $CD_{p,5}$ values that are available in the SciSciNet open data repository (Lin et al., 2023b), which features pre-calculated $CD_{p,5}$, along with the coauthor count k_p, and r_p values. Because computing $CD_{p,5}$ requires 5 years of post-publication citation data, in what follows we compare the disruptiveness of 18,644 research articles published in PNAS from 2011-2015. Notably, online-only *PNAS Plus* articles feature a different page numbering system, and so by inspecting this metadata for each article we identified 12.6% of the total sample as *PNAS Plus* articles.

We exploit this quasi-experimental setting in order to distinguish between the following behavioral (i) and statistical (ii & iii) mechanisms that could contribute to declines in CD:

(i) $N_j > N_i$: the main hypothesis for explaining the decline in CD put forward by Park et al. (2023) is that there have been fundamental shifts in scientific practice that have shifted away from disruptive science towards consolidating science. However, they do not eliminate the possibility that N_i is growing faster than N_j, on average, due to behavioral shifts affecting scholarly citation practice. Hence, increasing rates of N_j relative to N_i, may follow from a number of competing practical mechanisms, which they do discuss, but do not distinguish.

(ii) Statistical 'large N' convergence of CD: The distribution of CD is extremely concentrated around the centroid value of 0 – see **Fig. 2**(a.b). Hence, it is possible that $(N_i - N_j) \to 0$ as $c(t) \gg 1$ and $r(t) \gg 1$ increase over time, representing a statistical limit associated with increasing network density. Note that this candidate mechanism is reflected by the numerator of CD in Eq. (1).

(iii) Citation inflation: as $R_k \gg 1$, CD converges to 0 (since the numerator of CD is bounded). Unlike (ii), the source of this statistical mechanism is in the denominator of Eq. (1). Such convergence would explain the very small variance in the $CD_{p,5}$ distribution, which is extremely leptokurtic – see **Fig. 2**(c).

Park et al. (2023) attribute the observed decline in CD to shifting balance of disruptive innovation captured by mechanism (i). However, they do not rule out mechanisms (ii) or (iii), which are not related to the innovation capacity of the scientific enterprise, but instead reflect the susceptibility of the CD metric to statistical bias.

For this reason, we test the CI hypothesis using the absolute value, $|CD_{p,5}|$, which is more sensitive to the CI mechanism (iii) because positive adjustments to negative CD values and negative adjustments to positive CD values in response to shifts in positive definite metrics (i.e., $r_p > 0$) correspond to the same regression adjustment when using this absolute metric. This modification is similar in motivation to the alternative disruption metric employed by Lin et al. (2023a), who base their results on the sign of CD, which depends solely on the numerator $N_i - N_j$, and thus altogether avoids the measurement bias associated with mechanism (iii).

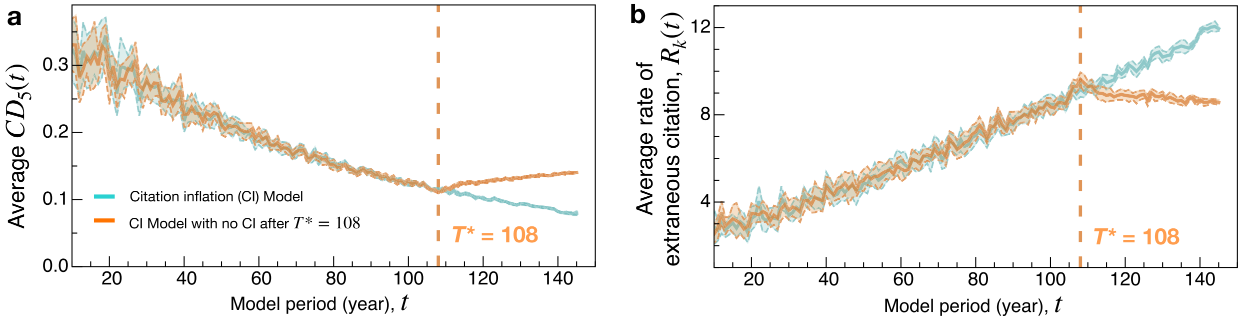

Fig. 3. Computational simulation of growing citation networks: after 'turning off' CI, the systematic decline in CD reverses. (a) The average $CD_5(t)$ calculated across 10 different computational realizations of (i) the standard CI model and (ii) the CI model with quenched reference list growth ($g_r = 0$) for $t \geq 108$. **(b)** Average rate of extraneous citation, $R_k(t)$, showing that $CD_5(t)$ converges to 0 because the denominator of the disruption index in Eq. (1) is unbounded as $r(t)$ grows.

3.3. Empirical multi-variate analysis of $CD_{p,5}$ – testing temporal and team-size relationships

We apply multi-variate regression in order to estimate the temporal trends in CD and the relationship between CD and team-size k_p. As in the previous section, we use pre-calculated data from SciSciNet (Lin et al., 2023b) for publications from the period 1995-2015 published by sources classified in the dataset as "Journal". To avoid data quality issues highlighted by Holst et al. (2024), we also implement the following thresholds: we exclude publications with reference list lengths outside the range $10 \leq r_p \leq 200$, which focuses on publications that are sufficiently embedded in the literature to ensure that they are likely products of dedicated research; we exclude team sizes outside the range $1 \leq k_p \leq 25$ to focus on the small and medium science regime analyzed by Wu et al. (2019); and we exclude publications outside the citation counts range $1 \leq c_{p,5} \leq 1000$, which avoids the definitional issue that $CD_p = 0$ for un-cited publications.

For the remaining set of publications that satisfy these thresholds, we then ranked the corresponding journals by their total publication volume over the 21-year period, and select the top 1000 most prominent journals. This journal exclusion produces just a 0.26% decrease in the sample size, resulting in 7,819,889 publications. Focusing on prominent journals facilitates defining a robust normalized disruption index,

$$\text{Norm}CD_{p,5,j,t} = \frac{CD_{p,5,j,t} - \overline{CD}_{j,t}}{\sigma[CD]_{j,t}} \ , \tag{2}$$

that is standardized at the year-journal level, where $\overline{CD}_{j,t}$ is the average CD value and $\sigma[CD]_{j,t}$ is the standard deviation calculated for publications from journal j in year t. See **Fig. 2**(c) for the frequency distribution $P(\text{Norm}CD_{p,5,j,t})$, which is also leptokurtic.

Importantly, this normalized disruption metric controls for year-specific factors such as journal publication modality (online, print, hybrid), as well as the characteristic value and variability of CD according to the discipline associated with j, etc. To ensure that results are not biased by extreme outliers, we exclude publications with $|\text{Norm}CD_{p,5,j,t}| \geq 5$, which corresponds to just a 0.66% decrease in the sample size. The resulting sample is comprised of 7,768,207 publications (comprising 99% of the original data sample), with the most productive (least productive) journal featuring 138,883 (respectively, 3371) publications over the 21-year period.

4. Results

4.1. Computational validation of the CI hypothesis

We generated and analyzed two ensembles of 10 random networks each. We report average values calculated across each ensemble so that our results are not sensitive to the idiosyncratic fluctuations of individual synthetic networks generated via our Monte Carlo simulation model. Thus, average differences between the two ensembles derive only from the presence or absence of CI in the generation of the synthetic citation networks.

Fig. 3 confirms that both network ensembles are statistically identical for the first 107 periods, which is by construction. However, in the second ensemble we eliminate the effects of CI from the citation network for $t \geq T^* = 108$ by quenching the growth of $r(t)$ such that $r(t) = 34$ for $t \geq T^*$. This sudden halt represents a hypothetical scenario in which journals were to impose a hard cap on reference list lengths in order to temper CI. Such caps are not inconceivable, as *Nature* provides a soft policy that "articles typically have no more than 50 references" in their formatting guide for authors.

Fig. 3(a) compares the trend in the average $CD_5(t)$ calculated for these two scenarios. The decline in CD reproduces the magnitude of empirical decline reported by Park et al. (2023) for $t < T^*$. However, the two curves diverge for $t \geq T^*$, with the curve featuring quenched reference lists suddenly reversing course, thereby revealing the acute effect of CI on CD. Similarly, **Fig. 1**(b) tracks the growth of $R_k(t)$, showing that this quantity is extensive when $r(t)$ is growing, and intensive when $r(t)$ is constant. Moreover, **Fig. S1**(c-f) shows that the model reproduces the relationship $R_k(t) \propto r(t)$, and also shows that the proportionality is independent of the citation window (CW) used for calculating $CD_{CW}(t)$; see Petersen et al. (2024) for additional empirical validation based upon a different dataset. Accordingly, our generative network model demonstrates that CI controls the trends in $CD(t)$.

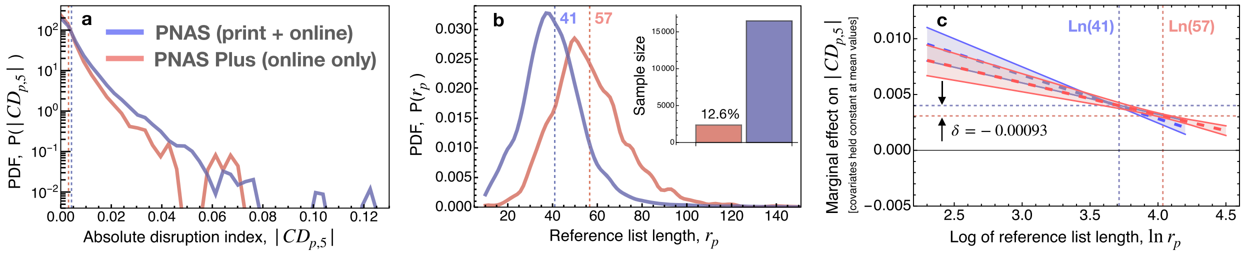

Fig. 4. Quasi-experimental test and validation of the CI hypothesis: counterfactual juxtaposition of research articles published in *PNAS* versus *PNAS Plus*. **(a)** Frequency distribution of the absolute disruption index, $|CD_{p,5}|$. **(b)** Frequency distribution of the number of references per paper, r_p. See **Fig. S3** for comparison of the two subsamples across a wider range of characteristics. Dashed vertical bars indicate the subsample means. **(c)** For both subsamples, the decline in is fully attributable to the variation in r_p such that the difference in average reference list lengths accounts for the entire, albeit small, difference in average $|CD_{p,5}|$.

4.2. Quasi-experimental validation of the CI hypothesis based upon $|CD_{p,5}|$

The difference in the average $|CD_{p,5}|$ value between the *PNAS* and *PNAS Plus* publication sets is incremental, with the *PNAS Plus* articles featuring smaller $|CD_{p,5}|$ values across the bulk of the sample distribution – see **Fig.** 4(a,b). In terms of relevant citation network characteristics related to the local subnetwork that defines CD_p, *PNAS Plus* articles differ primarily in terms of r_p, as they feature $100 \times (57 - 41)/41 = 39\%$ more references per article, on average. Otherwise, the two subsamples are nearly indistinguishable in terms of citation impact ($c_{p,5}$) and team size (k_p) – see **Fig. S3**.

We test the relationship between $|CD_{p,5}|$ and various covariates using the model specification employed in our companion study (Petersen et al., 2024), which is based upon disruption index values calculated from the Microsoft Academic Graph dataset (Sinha et al., 2015). Instead, here we use publicly available publication metadata from the *SciSciNet* open data repository (Lin et al., 2023b). We use the following multivariate linear regression model,

$$|CD_{p,5}| = b_t + b_k \ln k_p + b_r \ln r_p + b_c \ln c_{p,5} + \epsilon_t \ , \tag{3}$$

which accounts for team size and the most relevant scalar citation network quantities relating to CD. We estimate the parameters of the model using the STATA 13 package "xtreg fe" using publication-year fixed effects; each covariate enters in logarithm to temper the right-skew in the distribution of each variable. For the full list of parameter estimates see **Table S1**. As a demonstration of parameter estimate robustness and generalizability, we applied the same model to 6.9 million articles from the same period, 2011-2015, which shows consistent results across a larger range of journals – see **Table S2**.

In supports of the CI hypothesis, results show a negative relationship between $|CD_{p,5}|$ and r_p: $b_r = -0.0039$; $p = 0.002$; 95% CI $= [-0.0055 - 0.0023]$. Put in real terms, a paper with twice as many references ($2r_p$) has a $|CD_{p,5}|$ value that is $b_r \ln(2) = -0.002$ smaller than if it had r_p references. This scenario corresponds to a 0.6σ effect size, as the joint standard deviation across both PNAS subsamples is $\sigma[|CD_{p,5}|] = 0.0065$. Notably, the sign, magnitude, statistical significance level of b_r is consistent with the analog coefficient reported by Leahey et al. (2023). The relationship between $|CD_{p,5}|$ and k_p are not robust in sign, which is likely attributable to the small effect size compounded by the non-linear increasing relationship between k_p and r_p over time (Petersen et al., 2024), which we address in the following section.

The counterfactual design facilitates estimating the differences in $|CD_{p,5}|$ between the *PNAS* and *PNAS Plus* deriving solely from the differences in r_p. Our results show that 100% of the difference in the average $|CD_{p,5}|$ between the two journal subsets are explained by δ, the difference in the average r_p across the two subsets – see **Fig.** 4(c). Hence, these empirical results definitively demonstrate that a significant portion of variation in CD is attributable to variation in r_p. For this reason, the main results reported in Park et al. (2023) survived their robustness checks, e.g. the random rewiring they employed conserves r_p, and Extended Table 1 and Supplementary Table 3 show that they did not include r_p, c_p or k_p as publication-level covariates of CD. Since CD is a citation-network based indicator, other cited- and citing-document variables that explain citation counts (e.g., document, citation context, author and journal-oriented features) (Tahamtan & Bornmann, 2018) are needed to explain the remaining variation in CD. One limitation of such factor analysis methodologies is the difficulty in defining and measuring publication quality, which is an omitted variable in most studies, and limits the conclusiveness of estimated relationships.

4.3. Empirical re-analysis of 7.8 million publications from 1995-2015 – testing for CD correlations with t and k

There is considerable disagreement emerging from research analyzing the relationships between CD and various other factors. For example, Wu et al. (2019) mainly rely on descriptive methods to establish a negative relationship between CD_p and the team size, k_p. Instead, Petersen et al. (2024) and Leahey et al. (2023) employ multivariate regression and report a positive relationship, and no relationship between CD_p and k_p, respectively. One reason for the discrepancy emerging in the literature is a lack of consistency in the data and methodological specifications.

Hence, in this section we re-analyze publication-level temporal trends (Park et al., 2023) and team-size trends (Wu et al., 2019) in $CD_{p,5}$ using publicly available citation network data from *SciSciNet* (Lin et al., 2023b). We restrict our analysis to publications that feature explicit signatures of research outcomes – namely, those with sufficiently large r_p that we can be confident that they

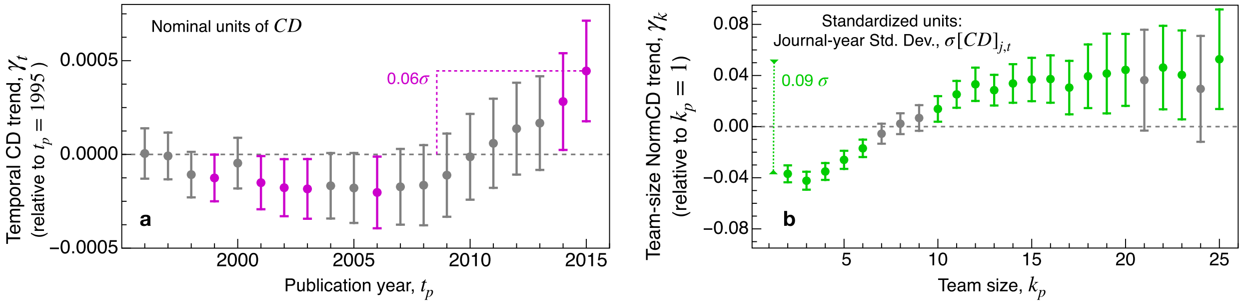

Fig. 5. Non-linear temporal and team-size trends in CD after controlling for CI. Marginal effects produced by multivariable regression that control for r_p and c_p (CI), increasing team sizes (k_p), and tendency for larger teams to produce longer papers with longer reference lists ($k_p \times t$). **(a)** Results indicate that disruptive science has incrementally increased since 2006 – which is consistent with three independent re-analyses reported by Bentley et al. (2023); Holst et al. (2024); Macher et al. (2024). The magnitude of the effect size (0.06σ) is relatively small. **(b)** In contrast to results reported by Wu et al. (2019), our results indicate that large teams (incrementally) disrupt and small teams (incrementally) develop science. The magnitude of the effect size (0.09σ) is inconsequential in terms of team science policy guidance and team assembly strategy. Shown are factor variable point estimates with 95% confidence intervals; Gray error bars are not statistically deviant from the baseline level indicated by the horizontal dashed line ($p > 0.05$). See **Tables S3 & S4** for the full list of model parameter estimates.

are not editorials, commentaries, book reviews or other non-research based content that may be misclassified as dedicated research. This selection also excludes publications featuring substantial missing network data (Macher et al., 2024), since these data quality issues effectively reduce r_p, and consequently give rise to spuriously large $\pm CD$; this selection also avoids the issue deriving from the surprisingly frequent singularity identified by Holst et al. (2024) whereby papers with $r_p = 0$ generate $CD_p = 1$. As such, we focus on the components of the citation network that are both conceivably and consequentially disruptive, in line with the originator's definition of disruption representing a form of breakthrough innovation (Christensen et al., 2015).

We analyze the temporal trend in CD using the following model specification,

$$CD_{p,5,j,t} = b_j + b_k \ln k_p + b_{k2}(\ln k_p)^2 + b_{k\times t}(\ln k_p \times t) + b_r \ln r_p + b_{r2}(\ln r_p)^2 + b_c \ln c_{p,5} + b_{c2}(\ln c_{p,5})^2 + \gamma_t + \epsilon_j \ , \quad (4)$$

which incorporates squared terms to account for non-linear relationships. For example, because $N_k \sim n(t)r(t)$ appears in the denominator of CD, a linear correction for r_p is likely insufficient. The coefficient $b_r = -0.0033$ (p-value < 0.001; 95% CI =[-0.0042, -0.0025]) is negative, reflecting the first-order residual impact of CI. For the full list of parameter estimates see **Table S3**.

Fig. 5(a) reports the trend in the factor variable γ_t, which captures year-specific trends that persist in spite of other publication-level controls. Note that the regression adjustment robustness checks in Supplementary Table 1 reported by Park et al. (2023) does not report any of the field-year and paper-level controls, and so it is not possible to validate our results according to their covariates; in particular, they not include the covariates r_p, k_p and $c_{p,5}$ in their model specification. Our reanalysis indicates that the residual trend in $CD(t)$ associated with time is at the level of noise, as the uptick in the regression-adjusted $CD(t)$ after 2008 corresponds to just a 0.06σ effect size relative to the baseline level in 1995.

In order to evaluate team-size trends, we leverage the journal-year normalized disruption index NormCD to estimate the standardized parameters of the model

$$\text{Norm}CD_{p,5,j,t} = b_t + \gamma_k + b_r \ln r_p + b_{r2}(\ln r_p)^2 + b_c \ln c_{p,5} + b_{c2}(\ln c_{p,5})^2 + \epsilon_t \ . \quad (5)$$

As such, coefficients are measured in units of $\sigma[CD]_{j,t}$, which facilitates assessing the relative magnitude of effect sizes. The interaction term represented by $(\ln k_p \times t)$ controls for the tendency of larger teams to produce longer papers with longer reference lists (Petersen et al., 2024). After controlling for temporal variation and CI, we find that CD increases (albeit weakly) with team size (for $k_p \in [3, 25]$) – which is consistent with a statistically significant and positive coefficient associated with $\ln k_p$ identified in our companion study (Petersen et al., 2024). As with the temporal trend, the net effect is at the level of noise, with the difference between $k_p = 2$ and $k_p = 25$ corresponding to just a 0.09σ effect size. These results are in disagreement with the results reported Wu et al. (2019). One source of discrepancy is the methodology, as their descriptive analysis does not account for multivariable interactions. Moreover, Wu et al. base their analysis upon differentials in the *percentile* values of $CD_{p,5}$, which obscures the relatively small magnitude of the effect size obtained for nominal CD values, which are extremely narrowly distributed around $CD = 0$ as illustrated in **Fig. 2**.

5. Discussion

A growing body of research seeks to relate CD_p to time-dependent covariates such as team size (Wu et al., 2019), novelty (Leahey et al., 2023), the geographic dispersion of team members (Lin et al., 2023a), and citation impact (Wang et al., 2023b) – all of which are quantities that have systematically increased over time. A common pattern among these studies is a result of the form: as X increases, CD decreases. However, this class of result follows naturally from the susceptibility of correlations between $X(t)$ and $CD(t)$ to (a) temporal biases associated with the secular growth of the scientific enterprise, and (b) temporal biases associated with increasing data completeness of the citation network data over time.

Data quality issues, deriving from missing citations and references, generate a fundamental source of error that Holst et al. (2024); Macher et al. (2024) show are responsible for the anomalous decline in $CD(t)$ reported by Park et al. (2023). In their comprehensive patent re-analysis, Macher et al. (2024) show that missing references at the beginning of the patent data artificially reduce r_p for early patents; upon correcting for their omission, which effectively increases r_p for those early patents, the negative trend in $CD(t)$ largely disappears. Similarly, in their combined publication and patent re-analysis, Holst et al. (2024) show that a significant source of systematic error follows from including items with $r_p = 0$ that generate $CD_p = 1$ outliers. They show that these anomalies tend to occur earlier in the publication and patent datasets. Upon correcting these issues, they also show that the negative trend in $CD(t)$ largely disappears. This second re-analysis also provides the full set of coefficients estimated in their regression adjustment analysis, which shows a negative correlation between CD_p and r_p – see β_1 in Table S1 reported by Holst et al. (2024). Indeed, these data quality issues give rise to the same net effect attributable to CI. Beyond data quality issues, another issue is the small effect size measuring correlations between CD and relevant covariates. Our re-analysis of temporal trends and team-size trends generate effect sizes at the 0.06σ and 0.09σ level, respectively; moreover, the directions of the trends are in disagreement with previous studies (Park et al., 2023; Wu et al., 2019).

To summarize, even in the absence of data quality issues, the CD index decreases over time due to two mechanisms unrelated to innovation – one behavioral, and the other structural. Importantly, as formulated, the disruption index does not account for confounding shifts in citation behavior (e.g. self-citation, impact factor boosting) that increase the rate of triadic closure measured by N_j in the numerator of CD. Thus, decreases in CD could follow from a number of competing mechanisms, some behavioral and reflecting a number of citation factors that evolved over time (Tahamtan & Bornmann, 2018), others statistical in nature. For this reason, alternative variants of the disruption index (Bornmann et al., 2020; Park et al., 2023) such as CD^{nok} that excludes N_k from Eq. (1) are also biased, because shifts in scholarly practice that manifest as network autocorrelations, such as self-citation and journal impact-factor boosting, increase the overall rate of consolidation (triadic closure) measured by the term N_j. The scientometrics community should be aware of this issue, which we believe is valuable to inform future research problem and metrics choices; we do stop short of recommending any specific alternative formulations of CD, however, in order to maintain neutrality on the issue of next steps. In order to facilitate the development of unbiased citation-network metrics, we do make available the ensemble of synthetic citation networks so that they can be used to test future citation-based indices for systematic bias (see Data Availability statement).

Instead, in this work we focused on testing the 'CI hypothesis' that underlies the structural mechanism highlighted above. Indeed, shifts in strategic behavior and normative practice are challenging to directly measure. For this reason, we confront this issue via computational simulation in our companion work (Petersen et al., 2024). In short, our mixed method approaches consistently demonstrate that CI causes the denominator of CD defined in Eq. (1) to systematically increase as reference lengths increase over time, which causes CD to converge to 0. According to its present definition, there is no clear way to correct for this dependence, since CD_p is non-linearly related to r_p via the factor N_k. This susceptibility is illustrated in **Fig. 1**, which shows how a publication (or patent) needs to only cite one highly-cited publication for N_k to increase to the extent that $CD \to 0$, independent of the difference $N_i - N_j$. The likelihood and magnitude of this one-off mechanism is increasing over time as a result of CI (Pan et al., 2018).

This issue is not merely a temporal bias, it also affects publications from the same publication cohort that have significantly different r_p. As a case example, we juxtaposed the disruptiveness of PNAS versus PNAS Plus articles published from 2011-2015, which differ primarily in their article lengths. Results show that nearly all of the difference in disruptiveness is attributable to the PNAS Plus articles having larger r_p on account of their extended online-only publication format. Hence, a significant amount of the variation in CD derives from variation in r_p, which could follow simply from journal-specific constraints on article lengths. By way of example, our analysis based upon the temporally-standardized disruption measure Norm$CD_{p,5}$ defined in Eq. (2) shows that the covariate with the largest effect size is r_p, which features a 0.14σ effect size for each unit change in $\ln r_p$ – see **Table S4**.

While in this work we do not propose specific modifications to the disruption metric that address the various measurement bias challenges we identified, we do recommend that citation metrics satisfy the following three conditions. First, in support of cross-temporal analysis, metrics should follow a stationary distribution over time, such as with normalized citation metrics that leverage the log-normal distribution of citation counts (Petersen et al., 2018; Radicchi et al., 2008). Second, metrics should be at most weakly sensitive to the secular growth of the academic enterprise, in particular CI deriving from increasing $r(t)$ and $n(t)$. And third, citation metrics should capture the consensus of the broader scientific community (as with citation counts) and not be entirely dependent on author choices (as with the selection of items included in a reference list).

We conclude with a hypothetical policy consideration regarding the management of the scientific publishing enterprise. As reference list lengths become longer, some journals now propose soft caps on the number of references allowed, which is more typical in letters journals such as *Nature*. While supporting acuity and conciseness, such caps also temper the effects of CI in research evaluation. Our computational simulation shows that journal policies that limit r_p could ameliorate the systematic decrease in $CD(t)$, and could simultaneously address related shortcomings in citation practice – such as, surgical self-citation by authors (Ibrahim et al., 2024; Ioannidis et al., 2019), large-scale citation rings (Evdaimon et al., 2024), institutional and national collectives (Qiu & Qiu, 2024; Tang et al., 2015), and journal impact-factor boosting (Ioannidis & Thombs, 2019; Martin, 2016).[1]

[1] Yet much deeper modifications to academic publishing process and culture are needed in order to fully stymie such citation trends. It is worth noting that the patent citation system involves patent examiners who mediate the annotation of prior art citations. Implementation of third-party reference list annotators in academic journals could reduce the phenomena of surgical self-citation, and points to the use-case of unbiased academic-AI examiners to serve this role in the academic peer-review process.

CRediT authorship contribution statement

Alexander Michael Petersen: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Felber J. Arroyave:** Writing – review & editing, Conceptualization. **Fabio Pammolli:** Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare that they have no competing financial interests.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.joi.2024.101605.

Data availability

Synthetic citation networks and code for analyzing them are available in the Dryad open data repository: DOI: 10.6071/M3G674. Data used for the multivariate regression analysis, including the disruption index $CD_{p,5}$, $c_{p,5}$, k_p and r_p, were obtained from the *SciSciNet* open data repository (Lin et al., 2023b).

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 1–3.
Barabasi, A. (2016). *Network science*. Cambridge University Press.
Bentley, R. A., Valverde, S., Borycz, J., Vidiella, B., Horne, B. D., Duran-Nebreda, S., & O'Brien, M. J. (2023). Is disruption decreasing, or is it accelerating? *Advances in Complex Systems*, *26*, Article 23500066.
Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020). Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. *Quantitative Science Studies*, *1*, 1242–1259.
Christensen, C., Raynor, M. E., & McDonald, R. (2015). What is disruptive innovation? *Harvard Business Review*, 3–11.
Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., et al. (2021). Advancing mathematics by guiding human intuition with AI. *Nature*, *600*, 70–74.
Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). Gpts are gpts: Labor market impact potential of llms. *Science*, *384*, 1306–1308.
Evdaimon, I., Ioannidis, J., Nikolentzos, G., Chatzianastasis, M., Panagopoulos, G., & Vazirgiannis, M. (2024). Metrics to detect small-scale and large-scale citation orchestration. Preprint, arXiv:2406.19219.
Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*, *63*, 791–817.
Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing* (p. 2116). NIH Public Access volume 2016.
Holst, V., Algaba, A., Tori, F., Wenmackers, S., & Ginis, V. (2024). Dataset artefacts are the hidden drivers of the declining disruptiveness in science. arXiv:2402.14583.
Huang, Y., Chen, L., & Zhang, L. (2020). Patent citation inflation: The phenomenon, its measurement, and relative indicators to temper its effects. *Journal of Informetrics*, *14*, Article 101015.
Ibrahim, H., Liu, F., Zaki, Y., & Rahwan, T. (2024). Google scholar is manipulatable. Preprint, arXiv:2402.04607.
Ioannidis, J. P., Baas, J., Klavans, R., & Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLoS Biology*, *17*, Article e3000384.
Ioannidis, J. P., & Thombs, B. D. (2019). A user's guide to inflated and manipulated impact factors. *European Journal of Clinical Investigation*, *49*, Article e13151.
Kozlov, M. (2023). 'Disruptive' science has declined - and no one knows why. *Nature*, *613*, 225.
Krapivsky, P. L., & Redner, S. (2005). Network growth by copying. *Physical Review E*, *71*, Article 036118.
Krenn, M., & Zeilinger, A. (2020). Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences*, *117*, 1910–1916.
Kristiansen, G. (2008). Style-shifting and shifting styles: A socio-cognitive approach to lectal variation. *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*, 4588.
Leahey, E., Lee, J., & Funk, R. J. (2023). What types of novelty are most disruptive? *American Sociological Review*, *88*, 562–597.
Leibel, C., & Bornmann, L. (2024). What do we know about the disruption indicator in scientometrics? An overview of the literature. *Scientometrics*, *129*, 601–639.
Leydesdorff, L., Tekles, A., & Bornmann, L. (2021). A proposal to revise the disruption index. *El Profesional de la Información*, *30*.
Li, H., Tessone, C. J., & Zeng, A. (2024). Productive scientists are associated with lower disruption in scientific publishing. *Proceedings of the National Academy of Sciences*, *121*, Article e2322462121.
Lin, Y., Frey, C. B., & Wu, L. (2023a). Remote collaboration fuses fewer breakthrough ideas. *Nature*, *623*, 987–991.
Lin, Z., Yin, Y., Liu, L., & Wang, D. (2023b). SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data*, *10*, 315.
Macher, J. T., Rutzer, C., & Weder, R. (2024). Is there a secular decline in disruptive patents? Correcting for measurement bias. *Research Policy*, *53*, Article 104992.
Martin, B. R. (2016). Editors' JIF-boosting stratagems – Which are appropriate and which not? *Research Policy*, *45*, 1–7.
Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., & Cubuk, E. D. (2023). Scaling deep learning for materials discovery. *Nature*, *624*, 80–85.
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*, 656–678.
Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, *613*, 138–144.
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*, 734–745.
Petersen, A. M. (2022). Evolution of recombinant biomedical innovation quantified via billions of distinct article-level MeSH keyword combinations. *Advances in Complex Systems*, *24*, Article 2150016.
Petersen, A. M., Arroyave, F., & Pammolli, F. (2024). The disruption index is biased by citation inflation. *Quantitative Science Studies*, *5*. https://doi.org/10.1162/qss_a_00333.

Petersen, A. M., Fortunato, S., Pan, R. K., Kaski, K., Penner, O., Rungi, A., Riccaboni, M., Stanley, H. E., & Pammolli, F. (2014). Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 15316–15321.

Petersen, A. M., Pan, R. K., Pammolli, F., & Fortunato, S. (2018). Methods to account for citation inflation in research evaluation. *Research Policy*, *48*, 1855–1865.

Qiu, S., Steinwender, C., & Azoulay, P. (2024). Paper Tiger? Chinese Science and Home Bias in Citations. National Bureau of Economic Research WP 32468; e-print:. https://doi.org/10.3386/w32468.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 17268–17272.

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

Schekman, R. (2010). Creating a new option for online-only research articles: PNAS plus. *Proceedings of the National Academy of Sciences*, *107*, 15309.

Schumpeter, J. A. (1942). *Capitalism, socialism and democracy*. New York, NY: Harper & Brothers.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., & Wang, K. (2015). An overview of Microsoft academic service (MAS) and applications. In *Proceedings of the 24th international conference on world wide web* (pp. 243–246).

Tahamtan, I., & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics*, *12*, 203–216.

Tang, L., Shapira, P., & Youtie, J. (2015). Is there a clubbing effect underlying Chinese research citation increases? *The Journal of the Association for Information Science and Technology*, *66*, 1923–1932.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, *571*, 95–98.

Verma, I. M. (2012). PNAS plus: Refining a successful experiment. *Proceedings of the National Academy of Sciences*, *109*, 13469.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. (2023a). Scientific discovery in the age of artificial intelligence. *Nature*, *620*, 47–60.

Wang, R., Zhou, Y., & Zeng, A. (2023b). Evaluating scientists by citation and disruption of their representative works. *Scientometrics*, *128*, 1689–1710.

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

Wu, S., & Wu, Q. (2019). A confusing definition of disruption. SocArXiv e-print:. https://doi.org/10.31235/osf.io/d3wpk.

Yang, D., Pavlidis, I., & Petersen, A. M. (2023). Biomedical convergence facilitated by the emergence of technological and informatic capabilities. *Advances in Complex Systems*, *26*, Article 2350003.