nature portfolio

https://doi.org/10.1038/s41586-022-05543-x

Supplementary information

Papers and patents are becoming less disruptive over time

In the format provided by the authors and unedited

Supplementary information for

"Papers and patents are becoming less disruptive over time"

Michael Park, Erin Leahey, and Russell J. Funk

Supplementary information

1. Changes in the diversity and novelty of word use over time

We examined changes in the diversity and novelty of word use in papers and patents over time. Our rationale for these analyses is that decreases in disruption should be associated with decreases in the diversity of words used in science and technology. Disruptive discoveries and inventions create departures from the status quo, rendering their predecessors less useful. While this pattern alone may have the effect of reducing the diversity of words used, disruptive discoveries and inventions are also likely to introduce new words or combinations of words; part of the way that disruptive discoveries and inventions render their predecessors less useful is by introducing ideas that are more useful than those that came before, which are likely to require new words or combinations of existing words to describe. Taken together with the long memory of language (i.e., even obsolete words are still occasionally used[1]), we therefore anticipate a positive association between disruption and the diversity and combinatorial novelty of word use by scientists and inventors. Thus, to the extent that our observations of decreasing disruption hold, we should see a decline in the diversity of words and novelty of their pairings over time.

To evaluate for such changes, we pulled all titles and abstracts for papers and patents in our sample from Web of Science and Patents View. For titles, there was very little missing data in either Web of Science or Patents View, with titles absent in fewer than 0.01% of cases in both the former and the latter. For abstracts, Patents View also provides highly complete coverage, with only 0.32% of cases missing. Web of Science has less robust coverage of abstracts before the early 1990s; from 1945-1991, only 4.45% of papers in our sample include abstracts. Coverage is much better in later years; from 1992-2010, abstracts are included for 90.85% of papers. We therefore limit our analyses of abstract data from WoS to the 1992-2010 period.

After extracting paper and patent titles and abstracts, we completed a series of preprocess-

ing steps using spaCy, an open-source, state-of-the-art Python package for natural language processing. To begin, we tokenized each title and abstract. From the resulting lists of tokens, we then excluded those that were tagged by spaCy as stop words, tokens consisting only of digits or punctuation, and tokens that were shorter than three characters or longer than 250 characters in length. Next, we converted all remaining tokens to their lemmatized form and converted all letters to lowercase. Finally, we aggregated the resulting lists of tokens to the subfield $\times$ year level, separately for papers and patents and for titles and abstracts.

We evaluate changes in the diversity of words used over time by computing, for each subfield $\times$ year, the *type-token ratio*, defined as the ratio of unique words to total words. We compute this measure separately for papers and patents and for titles and abstracts, at the level of the Web of Science research area (for papers) and NBER technology category (for patents). More specifically, for each field (i.e., research area or technology category) and each year, we divide the number of unique words appearing in titles by the total number of words appearing in titles (Figure 3a and 3d). We repeat this step for the number of unique and total words appearing in abstracts (Extended Data Figure 3). The measure attains its theoretical maximum when every word is used exactly once. Thus, higher values indicate greater diversity. We find similar results when measuring unique/total words using normalized entropy.

We followed a similar approach to measure the novelty of word pairs. For each subfield $\times$ year, we identify all pairwise combinations of words appearing in the titles of papers and patents. We then remove from that set all pairwise combinations of words in the titles of papers and patents published in the focal subfield in all previous years, which yields our count of new combinations. We then obtain our final measure by dividing this count of new combinations by the total number of word combinations made in the titles of papers and patents in the focal subfield and focal year (Figure 3b and 3e). Higher values indicate greater combinatorial novelty in word use.

2. Changes in word use over time

We also examined changes in the specific words used in papers and patents over time. Our rationale for these analyses is that the changes we observe in the CD index are likely to coincide with changes in approaches to discovery and invention, particularly the orientation of scientists and inventors towards prior knowledge. For example, to the extent that disruption is decreasing over time, it seems plausible that we will also observe decreases in words indicating the creation, discovery, or perception of new things. Similarly, it is also plausible that we will observe concomitant increases in the use of words that are more indicative of improvement, application, or assessment of existing things, which, consistent with the notion of consolidation, may reinforce existing streams of knowledge.

To evaluate changes in the use of specific words over time, we followed an approach similar to that described in our analyses of unique/total words, using similar samples of papers and patents and preprocessing steps. To simplify the presentation, we limit our attention to words appearing in paper and patent titles, for which, as noted previously, we have more complete data. However, the patterns we report below are also observable in analyses using paper and patent abstracts. Prior work has studied word frequencies in paper and patent titles extensively, and they are generally thought to provide a good window into the nature of science and technology[2]. For the present analyses, during preprocessing, we also assigned a part of speech tag to each lemma, after which we extracted all nouns, verbs, adjectives, and adverbs, which we anticipated would provide the most meaningful insights. At this stage, our data consisted of counts of lemmas by part of speech appearing in the titles of sample papers and patents. To facilitate analysis, we subsequently reshaped the data in a long-panel format, separately for papers and patents, where each row was uniquely identified by a document id $\times$ part of speech $\times$ token.

We then examined changes in the top 10 most frequently used words in paper and patent titles by decade. For patents, we present these word frequencies for the years 1980 and 2010;

for papers, our time series is longer, and therefore we present frequencies for 1950 and 2010. Analyses (available upon request from the authors) that include additional years (e.g., for each decade) yielded consistent results.

To simplify the presentation and conserve space, we focus our reporting on the results for verbs, which also generally yielded more noteworthy patterns (the most frequent nouns were often topical in nature; the most frequent adverbs and adjectives tended to be general/stop words).

We find evidence of a qualitative shift in word use that is consistent with our quantitative findings on the decline in disruptive activity using the CD index. Figure 3 shows the most common verbs in paper (Figure 3c) and patent titles (Figure 3f) in the first and last decade of each sample. In earlier decades, for example, verbs evoking creation (e.g., "produce", "form", "prepare", "make"), discovery (e.g., "determine", "report"), and perception (e.g., "measure") are prevalent in both paper and patent titles. In later decades, these verbs are almost completely displaced by those more evocative of the improvement (e.g., "improve", "enhance", "increase"), application (e.g., "use", "include"), or assessment (e.g., "associate", "mediate", "relate") of existing scientific and technological knowledge and artifacts. Thus, we observe a decrease in the use of verbs indicative of processes of disruption and a simultaneous increase in the use of verbs indicative of processes of consolidation.

These results are especially noteworthy when recalling that they are based on raw data, with no adjustment or transformation other than basic text preprocessing. Overall, then, the patterns offer compelling support for the findings we observe on the changing nature of science and technology using the CD index.

3. Disruptiveness and the growth of knowledge

Extended Data Figure 9 plots the number of new papers published and archived by WoS (panel a) and utility patents granted by the United States Patent and Trademark Office each year (panel b). The figure shows that there has been a sharp and consistent increase in the number of new papers and patents. The rate of new papers and patents added year by year seems to be accelerating, expanding the existing stock of knowledge rapidly, and thereby potentially placing a "burden of knowledge" on scientists and inventors.

Models in Supplementary Table 2 evaluate the relationship between the growth of knowledge and disruptiveness. Models 1 and 4 proxy for new knowledge based on the number of new works (patents or papers) produced in the focal year. Models 2 and 5 proxy for new knowledge based on the number of new works (patents or papers) produced in the five most recent years. Models 3 and 6 proxy for new knowledge based on the number of new works (patents or papers) produced in the ten most recent years. We find divergent results for papers and patents; for the former, there is a negative association between new knowledge and CD_5; for the latter, the association is positive. This divergent pattern motivates our subsequent analyses (Extended Data Table 1), on the use of prior knowledge and disruptiveness.

Supplementary Table 1: Regression models of trends in disruptiveness adjusted for publication, citation, and authorship practices

	Sample: Web of Science								Sample: Patents View							
	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)	
	b	se	b	se	b	se	b	se	b	se	b	se	b	se	b	se
Year=1946	-0.01***	0.00	-0.01***	0.00	-0.01***	0.00	-0.01***	0.00								
Year=1947	-0.02***	0.00	-0.02***	0.00	-0.02***	0.00	-0.02***	0.00								
Year=1948	-0.04***	0.00	-0.03***	0.00	-0.03***	0.00	-0.03***	0.00								
Year=1949	-0.07***	0.00	-0.07***	0.00	-0.07***	0.00	-0.07***	0.00								
Year=1950	-0.10***	0.00	-0.10***	0.00	-0.10***	0.00	-0.10***	0.00								
Year=1951	-0.11***	0.00	-0.11***	0.00	-0.11***	0.00	-0.11***	0.00								
Year=1952	-0.14***	0.00	-0.13***	0.00	-0.13***	0.00	-0.13***	0.00								
Year=1953	-0.15***	0.00	-0.15***	0.00	-0.14***	0.00	-0.15***	0.00								
Year=1954	-0.16***	0.00	-0.15***	0.00	-0.15***	0.00	-0.15***	0.00								
Year=1955	-0.14***	0.00	-0.13***	0.00	-0.13***	0.00	-0.14***	0.00								
Year=1956	-0.14***	0.00	-0.13***	0.00	-0.13***	0.00	-0.14***	0.00								
Year=1957	-0.14***	0.00	-0.14***	0.00	-0.13***	0.00	-0.14***	0.00								
Year=1958	-0.15***	0.00	-0.15***	0.00	-0.14***	0.00	-0.15***	0.00								
Year=1959	-0.16***	0.00	-0.16***	0.00	-0.15***	0.00	-0.17***	0.00								
Year=1960	-0.17***	0.00	-0.16***	0.00	-0.16***	0.00	-0.17***	0.00								
Year=1961	-0.16***	0.00	-0.15***	0.00	-0.15***	0.00	-0.16***	0.00								
Year=1962	-0.18***	0.00	-0.17***	0.00	-0.16***	0.00	-0.18***	0.00								
Year=1963	-0.18***	0.00	-0.17***	0.00	-0.17***	0.00	-0.18***	0.00								
Year=1964	-0.17***	0.00	-0.17***	0.00	-0.16***	0.00	-0.18***	0.00								
Year=1965	-0.18***	0.00	-0.18***	0.00	-0.17***	0.00	-0.19***	0.00								
Year=1966	-0.18***	0.00	-0.18***	0.00	-0.17***	0.00	-0.19***	0.00								
Year=1967	-0.19***	0.00	-0.19***	0.00	-0.18***	0.00	-0.20***	0.00								
Year=1968	-0.20***	0.00	-0.20***	0.00	-0.18***	0.00	-0.21***	0.00								
Year=1969	-0.20***	0.00	-0.20***	0.00	-0.19***	0.00	-0.22***	0.00								
Year=1970	-0.21***	0.00	-0.21***	0.00	-0.20***	0.00	-0.22***	0.00								
Year=1971	-0.22***	0.00	-0.21***	0.00	-0.20***	0.00	-0.23***	0.00								
Year=1972	-0.22***	0.00	-0.22***	0.00	-0.21***	0.00	-0.24***	0.00								
Year=1973	-0.22***	0.00	-0.22***	0.00	-0.20***	0.00	-0.24***	0.00								
Year=1974	-0.21***	0.00	-0.20***	0.00	-0.19***	0.00	-0.23***	0.00								
Year=1975	-0.22***	0.00	-0.22***	0.00	-0.21***	0.00	-0.25***	0.00								
Year=1976	-0.22***	0.00	-0.22***	0.00	-0.21***	0.00	-0.25***	0.00								
Year=1977	-0.20***	0.00	-0.20***	0.00	-0.19***	0.00	-0.23***	0.00								
Year=1978	-0.19***	0.00	-0.19***	0.00	-0.18***	0.00	-0.22***	0.00								
Year=1979	-0.21***	0.00	-0.21***	0.00	-0.19***	0.00	-0.24***	0.00								
Year=1980	-0.21***	0.00	-0.21***	0.00	-0.19***	0.00	-0.25***	0.00								
Year=1981	-0.21***	0.00	-0.21***	0.00	-0.19***	0.00	-0.25***	0.00	-0.05***	0.00	-0.05***	0.00	-0.05***	0.00	-0.05***	0.00
Year=1982	-0.22***	0.00	-0.22***	0.00	-0.20***	0.00	-0.26***	0.00	-0.09***	0.00	-0.09***	0.00	-0.09***	0.00	-0.09***	0.00
Year=1983	-0.22***	0.00	-0.22***	0.00	-0.20***	0.00	-0.26***	0.00	-0.13***	0.00	-0.13***	0.00	-0.12***	0.00	-0.13***	0.00
Year=1984	-0.23***	0.00	-0.23***	0.00	-0.21***	0.00	-0.27***	0.00	-0.14***	0.00	-0.14***	0.00	-0.14***	0.00	-0.14***	0.00
Year=1985	-0.24***	0.00	-0.24***	0.00	-0.22***	0.00	-0.29***	0.00	-0.16***	0.00	-0.16***	0.00	-0.16***	0.00	-0.16***	0.00
Year=1986	-0.22***	0.00	-0.22***	0.00	-0.20***	0.00	-0.27***	0.00	-0.18***	0.00	-0.18***	0.00	-0.18***	0.00	-0.18***	0.00
Year=1987	-0.22***	0.00	-0.22***	0.00	-0.19***	0.00	-0.27***	0.00	-0.19***	0.00	-0.19***	0.00	-0.19***	0.00	-0.20***	0.00
Year=1988	-0.22***	0.00	-0.22***	0.00	-0.19***	0.00	-0.27***	0.00	-0.22***	0.00	-0.22***	0.00	-0.21***	0.00	-0.22***	0.00
Year=1989	-0.21***	0.00	-0.21***	0.00	-0.19***	0.00	-0.27***	0.00	-0.23***	0.00	-0.23***	0.00	-0.23***	0.00	-0.24***	0.00
Year=1990	-0.21***	0.00	-0.21***	0.00	-0.19***	0.00	-0.27***	0.00	-0.24***	0.00	-0.24***	0.00	-0.23***	0.00	-0.25***	0.00
Year=1991	-0.21***	0.00	-0.21***	0.00	-0.18***	0.00	-0.27***	0.00	-0.25***	0.00	-0.25***	0.00	-0.25***	0.00	-0.26***	0.00
Year=1992	-0.21***	0.00	-0.21***	0.00	-0.19***	0.00	-0.28***	0.00	-0.26***	0.00	-0.26***	0.00	-0.26***	0.00	-0.28***	0.00
Year=1993	-0.21***	0.00	-0.21***	0.00	-0.18***	0.00	-0.29***	0.00	-0.27***	0.00	-0.27***	0.00	-0.27***	0.00	-0.29***	0.00
Year=1994	-0.22***	0.00	-0.22***	0.00	-0.19***	0.00	-0.29***	0.00	-0.29***	0.00	-0.28***	0.00	-0.28***	0.00	-0.30***	0.00
Year=1995	-0.22***	0.00	-0.22***	0.00	-0.19***	0.00	-0.30***	0.00	-0.29***	0.00	-0.29***	0.00	-0.28***	0.00	-0.30***	0.00
Year=1996	-0.26***	0.00	-0.26***	0.00	-0.23***	0.00	-0.34***	0.00	-0.29***	0.00	-0.29***	0.00	-0.29***	0.00	-0.31***	0.00
Year=1997	-0.26***	0.00	-0.26***	0.00	-0.23***	0.00	-0.35***	0.00	-0.30***	0.00	-0.30***	0.00	-0.29***	0.00	-0.31***	0.00
Year=1998	-0.26***	0.00	-0.26***	0.00	-0.23***	0.00	-0.35***	0.00	-0.30***	0.00	-0.30***	0.00	-0.29***	0.00	-0.32***	0.00
Year=1999	-0.26***	0.00	-0.26***	0.00	-0.23***	0.00	-0.35***	0.00	-0.30***	0.00	-0.30***	0.00	-0.29***	0.00	-0.33***	0.00
Year=2000	-0.26***	0.00	-0.26***	0.00	-0.23***	0.00	-0.36***	0.00	-0.31***	0.00	-0.31***	0.00	-0.31***	0.00	-0.34***	0.00
Year=2001	-0.26***	0.00	-0.26***	0.00	-0.23***	0.00	-0.36***	0.00	-0.32***	0.00	-0.32***	0.00	-0.31***	0.00	-0.35***	0.00
Year=2002	-0.26***	0.00	-0.26***	0.00	-0.23***	0.00	-0.36***	0.00	-0.33***	0.00	-0.33***	0.00	-0.32***	0.00	-0.36***	0.00
Year=2003	-0.26***	0.00	-0.26***	0.00	-0.23***	0.00	-0.37***	0.00	-0.34***	0.00	-0.34***	0.00	-0.33***	0.00	-0.37***	0.00
Year=2004	-0.27***	0.00	-0.27***	0.00	-0.23***	0.00	-0.38***	0.00	-0.34***	0.00	-0.34***	0.00	-0.33***	0.00	-0.37***	0.00
Year=2005	-0.27***	0.00	-0.27***	0.00	-0.23***	0.00	-0.38***	0.00	-0.34***	0.00	-0.35***	0.00	-0.33***	0.00	-0.37***	0.00
Year=2006	-0.27***	0.00	-0.27***	0.00	-0.23***	0.00	-0.39***	0.00	-0.35***	0.00	-0.35***	0.00	-0.33***	0.00	-0.38***	0.00
Year=2007	-0.27***	0.00	-0.27***	0.00	-0.23***	0.00	-0.39***	0.00	-0.34***	0.00	-0.34***	0.00	-0.33***	0.00	-0.38***	0.00
Year=2008	-0.27***	0.00	-0.27***	0.00	-0.23***	0.00	-0.40***	0.00	-0.34***	0.00	-0.34***	0.00	-0.32***	0.00	-0.37***	0.00
Year=2009	-0.27***	0.00	-0.28***	0.00	-0.23***	0.00	-0.40***	0.00	-0.33***	0.00	-0.33***	0.00	-0.32***	0.00	-0.37***	0.00
Year=2010	-0.27***	0.00	-0.28***	0.00	-0.22***	0.00	-0.41***	0.00	-0.33***	0.00	-0.33***	0.00	-0.31***	0.00	-0.37***	0.00
Constant	0.28***	0.00	0.28***	0.00	0.30***	0.00	0.23***	0.00	0.38***	0.00	0.38***	0.00	0.38***	0.00	0.32***	0.00
Subfield fixed effects	No		Yes		Yes		Yes		No		Yes		Yes		Yes	
Paper/patent-level controls	No		No		Yes		Yes		No		No		Yes		Yes	
Field × year-level controls	No		No		No		Yes		No		No		No		Yes	
N	22479429		22456096		22456096		22456096		2926923		2926923		2926923		2926923	
R2	0.03		0.07		0.10		0.11		0.08		0.09		0.09		0.10	
Wald tests of controls																
F					19102.01		28046.91						9198.56		2974.44	
d.f.					2.00		5.00						1.00		4.00	
p-value					0.00		0.00						0.00		0.00	

Notes: Estimates are from ordinary-least-squares regressions. Robust standard errors are shown in parentheses. Models 1 and 5 are the baseline models. Models 2 and 6 add subfield fixed effects. Models 3-4 (papers) and 7-8 (patents) add the field × year- and paper/patent-level controls, including adjustments for field × year level—*Number of new papers/patents*, *Mean number of papers/patents cited*, *Mean number of authors/inventors per paper/patent*—and paper/patent-level—*Number of papers/patents cited*, *Number of unlinked references*—characteristics. The reference categories for the year indicators are 1945 and 1980 for papers and patents, respectively. Each coefficient is tested against the null hypothesis of being equal to 0 using a two-sided t-test. We do not adjust for multiple hypothesis testing. The Wald tests, shown below the models including the control variables, test the null hypothesis that parameters for the control variables are simultaneously zero using a two-sided F-test. Rejection of the null hypothesis is evidence that the control variables add to the explanatory power of the model.

*p<0.1 **p<0.05; ***p<0.01

Supplementary Table 2: Regression models of disruptiveness and the growth of knowledge

	Sample: Patents View			Sample: Web of Science		
	(1)	(2)	(3)	(4)	(5)	(6)
Number of new works in the field during focal year (logged)	0.0036***			-0.0042***		
	(0.0004)			(0.0001)		
Number of new works in the field during past 5 years (logged)		0.0026***			-0.0019***	
		(0.0002)			(0.0001)	
Number of new works in the field during past 10 years (logged)			0.0033***			-0.0019***
			(0.0002)			(0.0001)
Constant	0.0291***	0.0064	-0.0095*	0.0519***	0.0479**	0.0491***
	(0.0030)	(0.0044)	(0.0051)	(0.0016)	(0.0016)	(0.0017)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Field fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	3434055	3434055	3434055	2.16e+07	2.16e+07	2.16e+07
R2	0.06	0.06	0.06	0.02	0.02	0.02

Notes: This table evaluates the relationship between the number of new works (papers or patents, a proxy for the growth of knowledge) and CD_5. Estimates are from ordinary-least-squares regressions. Each coefficient is tested against the null hypothesis of being equal to 0 using a two-sided t-test. We do not adjust for multiple hypothesis testing. Robust standard errors are shown in parentheses.
*p<0.1 **p<0.05; ***p<0.01

Supplementary Table 3: Summary statistics for variables used in regression analyses

WoS/Papers

Variable	N	N_{unique}	Mean	SD	Min	Max	Level of measurement
Use of existing knowledge							
Diversity of work cited	24615975	8811	0.98	0.01	0.88	1.00	Field $\times$ year
Ratio of self-citations to total work cited	21766010	5779	0.16	0.20	0.00	1.00	Publication
Mean age of work cited	21766010	74735	9.22	5.55	0.00	110.00	Publication
Dispersion in age of work cited	21766010	2511142	6.32	4.22	0.00	53.00	Publication
Controls							
Mean age of team members	24621907	22636	13.67	10.76	0.00	80.00	Publication
Mean number of prior works produced by team members	24656790	58843	35.70	57.23	0.00	2381.00	Publication
Outcomes							
CD_5 index	22479429	848779	0.04	0.22	-1.00	1.00	Publication
Fixed effects							
Publication year	24659076	66	–	–	–	–	Year
WoS extended subject	24622172	151	–	–	–	–	Field

Patents View/Patents

Variable	N	N_{unique}	Mean	SD	Min	Max	Level of measurement
Use of existing knowledge							
Diversity of work cited	3912211	1278	0.98	0.01	0.84	1.00	Field $\times$ year
Ratio of self-citations to total work cited	3486663	6122	0.07	0.18	0.00	1.00	Publication
Mean age of work cited	3486663	47948	7.67	4.15	-36.00	34.00	Publication
Dispersion in age of work cited	3486663	278012	3.35	2.49	0.00	19.00	Publication
Controls							
Mean age of team members	3911607	2116	5.60	5.93	0.00	34.00	Publication
Mean number of prior works produced by team members	3911607	8340	10.51	49.70	0.00	2852.00	Publication
Outcomes							
CD_5 index	3682052	58722	0.12	0.30	-1.00	1.00	Publication
Fixed effects							
Publication year	3912353	35	–	–	–	–	Year
NBER technology subcategory	3912281	39	–	–	–	–	Field

Supplementary references

1. Michel, J.-B. *et al.* Quantitative analysis of culture using millions of digitized books. *Science* **331,** 176–182 (2011).
2. Milojević, S. Quantifying the cognitive extent of science. *Journal of Informetrics* **9,** 962–973 (2015).