PNAS
www.pnas.org

Supplementary Information for
Slowed Canonical Progress in Large Fields of Science

Johan S. G. Chu, James A. Evans.

Johan Chu.
Email: johan.chu@kellogg.northwestern.edu

This PDF file includes:

Supplementary text
Figure S1
Tables S1 to S3

Supplementary Information Text

Time versus number of papers. A salient alternative explanation for the observed results could be that the passage of time is driving both the increase in number of papers and the concentrated citation patterns. Small, new fields may explore various avenues of inquiry until they coalesce around a fruitful canon. The establishment of a generative canon can then foster and support the growth of the field. Our analyses cannot rule out this explanation directly, although theoretically, we doubt that any initial canon could provide sole and sufficient basis for optimal exploration to a scientific field for many decades. Indeed, our main policy argument is that the canon fails to shift sufficiently when the field becomes large.

Nonetheless, we can examine whether the effects of field size persist when considering effects of field age. Table S1 shows regressions where the dependent variable is the rank correlation of the top-50 most-cited articles across subsequent years. The number of papers published in the focal year remains a significant predictor of higher rank correlation even when including calendar year in the regression, either as a continuous variable or with year dummies. A 10-fold increase in the number of published papers in a year leads to a 0.09 increase in the rank correlation coefficient, a similar effect size to that of 15 years of field maturation.

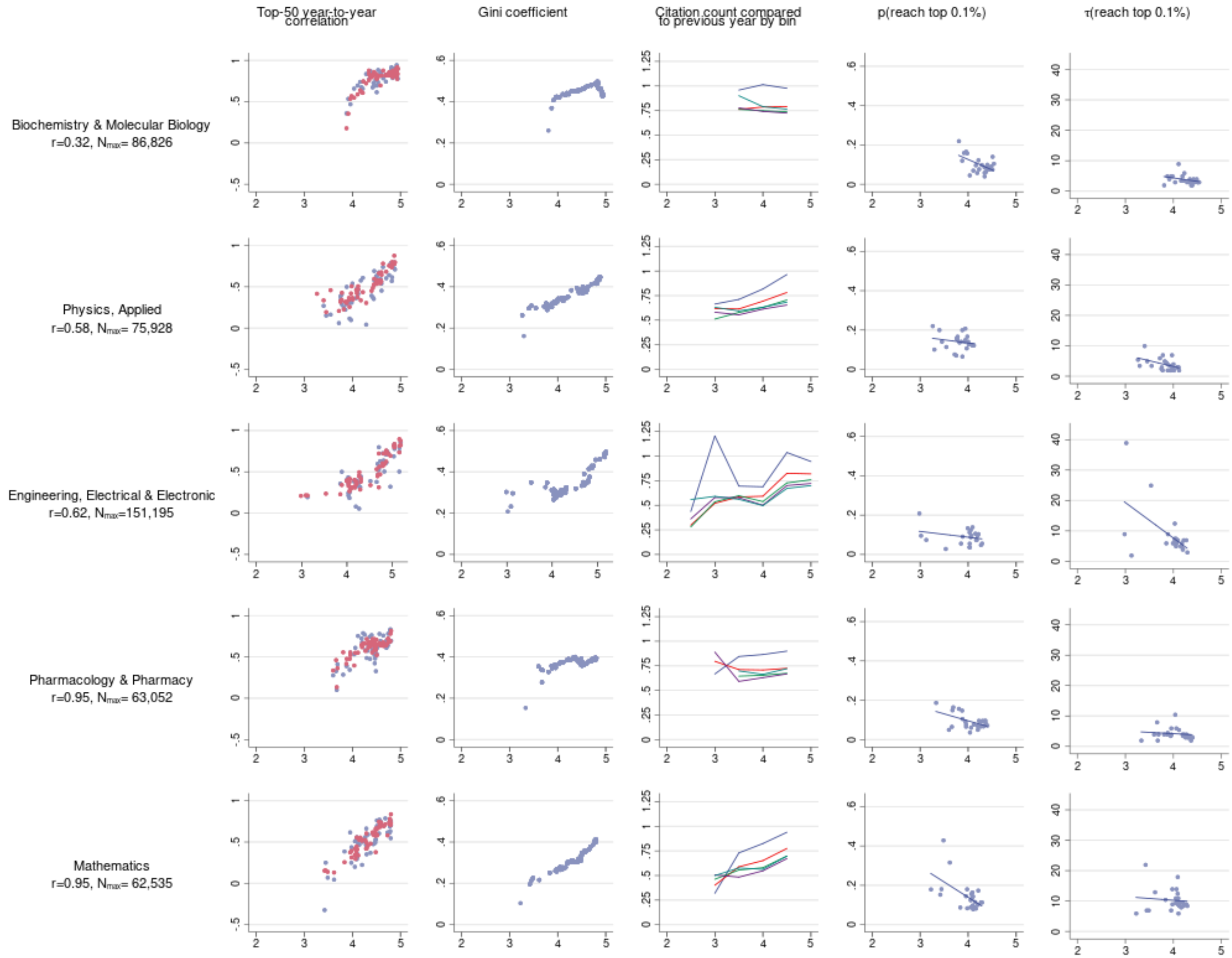
These results suggest that field size matters above and beyond any effect of the passage of time. But since field size tends to increase over time, one could question whether this effect size is believable given the degree of collinearity. Examining the 10 largest non-multidisciplinary fields (Fig. S1) suggests that the effect of field size is robust. Year and field size are not always very strongly correlated, with the correlation coefficient at the field level, dipping to 0.32 for Biochemistry & Molecular Biology, for example. Even when the correlation between year and field size are lower, citation patterns by size of field are similar. Larger fields have less churn in the most-cited list, the most-cited papers garner disproportionate shares of citations and have low decay in the number of citations year over year. New papers in larger fields have lower probability of ever becoming widely-cited, and when they do rise into the ranks of the most-cited, do so rapidly rather than through a cumulative attention-gathering process. While future research can further elucidate the relationships between field size, field age, and citation patterns, these analyses strongly suggest that size does matter.

Imprinting versus adaptation. Our main analysis demonstrates that as fields get larger, newly-published papers tend to cite the already most-cited more often. An intriguing question is whether long-time scholars are immune from these effects. Well-established, veteran scholars who were trained when a field was smaller may have entrained habits of reading new papers (rather than just scanning them and locating them relevant to established canon), possess the capacity to assimilate new information better, and may be better positioned to publish their papers without excessively relating their new work to established canon. If these advantages matter and persist, the shift in citation patterns we observed may be driven by cohort effects, where scholars entering the field in different periods cite differently, more than by period effects dependent on when papers are published.

Table S2 shows regression results for the largest field in our data, Electrical & Electronic Engineering. The dependent variables in the three models are the probability that a citation is to a top 0.1% most-cited, top 1% most-cited, and bottom 50% most-cited paper respectively. The regressions include fixed effects for authors.

We find a period effect for scholars above and beyond a cohort (or time of field entry) effect. Paper authors change their citation patterns depending on field size at time of paper publication. When the number of papers published in a field is 10 times larger, a citation by the same author is about 2% more likely to refer to a top 0.1% or 1% most-cited paper, and about 2% less likely to cite a bottom 50% paper.

The regressions shown in Table S2 do not rule out an initial imprinting or cohort effect. To estimate the size of such an effect, Table S3 shows regressions without author fixed effects and including field size at the time of the author's first published paper (N_e). The current-year size effect is much larger than the entry-year effect. A ten-fold increase in current field size leads to 3% and 6% increased propensities to cite top 0.1% most-cited papers and 1% papers respectively. The estimated effect of field size at entry is 16-30 times smaller. These regressions suggest even established, veteran scholars are forced to change their reading and citation patterns when fields grow large.



Top-50 year-to-year correlation
Gini coefficient
Citation count compared to previous year by bin
p(reach top 0.1%)
τ(reach top 0.1%)
Biochemistry & Molecular Biology
r=0.32, N_{max}= 86,826
Physics, Applied
r=0.58, N_{max}= 75,928
Engineering, Electrical & Electronic
r=0.62, N_{max}=151,195
Pharmacology & Pharmacy
r=0.95, N_{max}= 63,052
Mathematics
r=0.95, N_{max}= 62,535

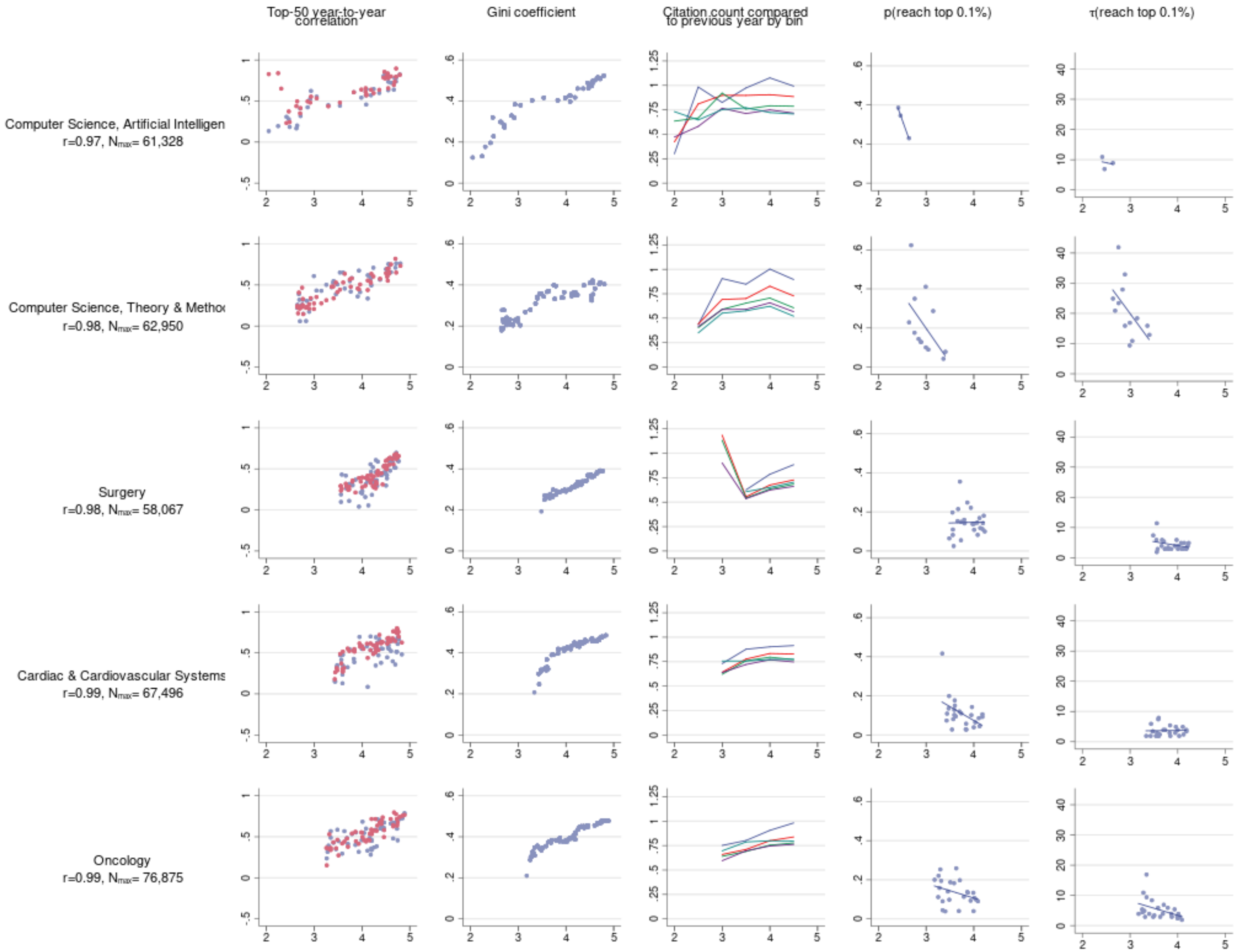

Top-50 year-to-year correlation
Gini coefficient
Citation count compared to previous year by bin
p(reach top 0.1%)
τ(reach top 0.1%)
Computer Science, Artificial Intelligen
r=0.97, N_max= 61,328
Computer Science, Theory & Methoc
r=0.98, N_max= 62,950
Surgery
r=0.98, N_max= 58,067
Cardiac & Cardiovascular Systems
r=0.99, N_max= 67,496
Oncology
r=0.99, N_max= 76,875

Fig. S1. Top-50 most-cited year-to-year correlation (blue dots: Spearman rank correlation, red dots: proportion of focal year top-50 remaining in subsequent year top-50); Gini coefficients of citation share; Citation decay rate; Probability (p, in %) of a paper ever reaching the top 0.1% of most-cited articles; Median number of years (τ) for a paper to reach the top 0.1% of most-cited articles for ten-largest non-multidisciplinary subjects. Subjects are ordered by their correlation coefficient r between logged (base 10) number of papers and year.

Table S1. Linear regression of rank correlation of the top-50 most-cited articles across subsequent years in a subject over size of subject in year (logged number of publications) and year.

DV: Rank correlation	Model 1 b/se	Model 2 b/se
$\log_{10}N$	0.091***	0.090***
	(0.01)	(0.01)
year	0.006***	
	(0.00)	
year dummies	No	Yes
subject fixed effects	Yes	Yes
constant	-12.113***	-0.238***
	(0.31)	(0.03)
R-sqr	0.431	0.440

* p<0.05, ** p<0.01, *** p<0.001

Table S2. Linear regressions of the probability a citation is to a top 0.1%/ top 1%/ bottom 50% most-cited paper with author fixed effects.

DV: p(cite)	top 0.1% b/se	top 1% b/se	bottom 50% b/se
log10 N	0.018***	0.021***	-0.022***
	(0.00)	(0.00)	(0.00)
year	-0.000***	-0.003***	0.005***
	(0.00)	(0.00)	(0.00)
author f.e.	Yes	Yes	Yes
constant	0.376***	5.313***	-10.427***
	(0.08)	(0.14)	(0.09)
R-sqr	0.000	0.002	0.020

* p<0.05, ** p<0.01, *** p<0.001

Table S3. Linear regressions of the probability a citation is to a top 0.1%/ top 1%/ bottom 50% most-cited paper without author fixed effects.

DV: p(cite)	top 0.1% b/se	top 0.1% b/se	top 0.1% b/se
log10 N	0.028***		0.030***
	(0.00)		(0.00)
log10 Ne		0.001***	0.001***
		(0.00)	(0.00)
year	0.001***	0.001***	0.001***
	(0.00)	(0.00)	(0.00)
author f.e.	No	No	No
constant	-1.318***	-2.701***	-1.214***
	(0.07)	(0.02)	(0.07)
R-sqr	0.009	0.008	0.009

	top 1% b/se	top 1% b/se	top 1% b/se
log10 N	0.065***		0.064***
	(0.00)		(0.00)
log10 Ne		0.004***	0.004***
		(0.00)	(0.00)
year	-0.000***	0.001***	-0.000***
	(0.00)	(0.00)	(0.00)
author f.e.	No	No	No
constant	0.375**	-2.693***	0.509***
	(0.12)	(0.04)	(0.12)
R-sqr	0.004	0.004	0.005

	bottom 50% b/se	bottom 50% b/se	bottom 50% b/se
log10 N	-0.039***		-0.047***
	(0.00)		(0.00)
log10 Ne		0.008***	0.007***
		(0.00)	(0.00)
year	-0.001***	-0.002***	-0.001***
	(0.00)	(0.00)	(0.00)
author f.e.	No	No	No
constant	1.459***	3.829***	1.470***
	(0.07)	(0.03)	(0.07)
R-sqr	0.010	0.020	0.021

* p<0.05, ** p<0.01, *** p<0.001