Quantifying the Evolution of Scientific Impact

Frontiners of Network Science Presentation

Brandon Rozek
rozekb@rpi.edu

Rensselaer Polytechnic Institute, Troy, NY, USA

November 2022

How do we identify promising scientists?

- **Publication:** Number or velocity of publications
- **Citations:** Total citations, h-index
- **Grants:** Number of grants awarded, total sum of award money
- Other awards, fellowships, etc.

Most Popular Metric: The H-Index

H-index = 10

Scientist has published **10** papers with **10** or more citations.

Main Issue

These measures monotonically increase throughout a researchers career.

Current Workaround

Instead of assessing in isolation, compare the scientist against their peers:

- Compare recent PhD graduates.
- NSF Career Panel only reviews applications from untenured research professors.

Alternatively, an assessment committee may compare a candidate against historical information.

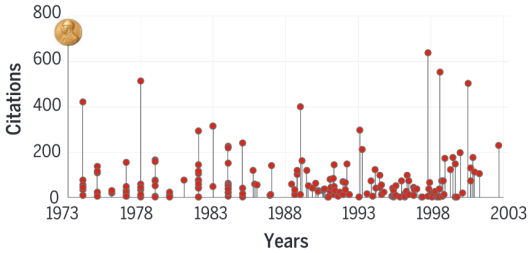
Goal

How do we create a measure of a scientists impact, that allows for comparison with any other scientist, regardless of seniority?

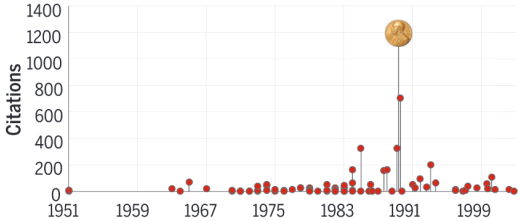
Prior Work

Sinatra, Wang, Deville, Song, and Barabasi
"Quantifying the evolution of individual scientific impact"
Science (2016).

Core Contribution 1: Impact is Randomly Distributed



Frank A. Wilczek
Physics Nobel,
2004



John B. Fenn
Chemistry Nobel,
2002

Core Contribution 2: Modeling Individual Impact

Modelled impact of a paper through a trivariate normal distribution based on the luck, a scientist's productivity, and their individual talent.

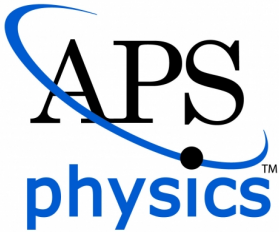
Outline

1. Prior Work: Dataset, Methods, & Findings
2. Current Work: Extensions, Dataset, & Findings
3. Future Work and Conclusion

Prior Work - Datasets

Dataset 1: APS Citation Dataset

The publication record of 236,884 physicists publishing in the Physical Review from 1893 to 2010 for a total of 450,000 publications.



Dataset 2: Google Scholar + Web of Science

The combination of 24,630 Google Scholar career profiles with Web of Science data, covering 514,896 publications in the natural and social sciences.

Dataset Filtering

- Publication record spans at least 20 years.
- Authored at least 10 total publications.
- Wrote at least 1 paper every 5 years

Graphics in the paper are from the initial dataset featuring 2887 physicists.

Prior Work - Methods

Issues in Citation Based Methods

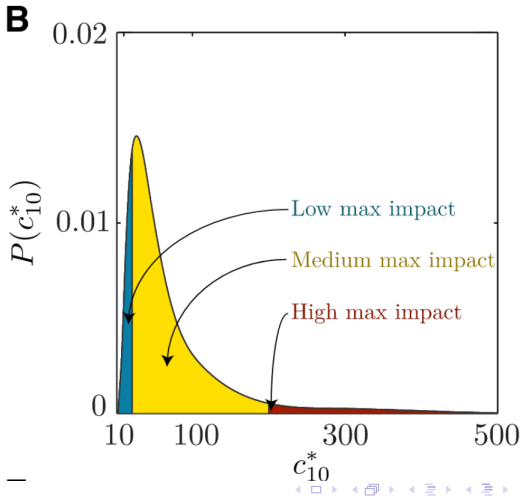
1. Citations follow different dynamics for different papers
2. Average number of citations changes over time
3. Citation count is subfield-dependent

Techniques to overcome these issues

1. For each paper, we calculate c_{10} which is the cumulative number of citations the paper received 10 years after its publication.
2. Normalizing c_{10} by the average $\langle c_{10} \rangle$ of papers published in the same year.

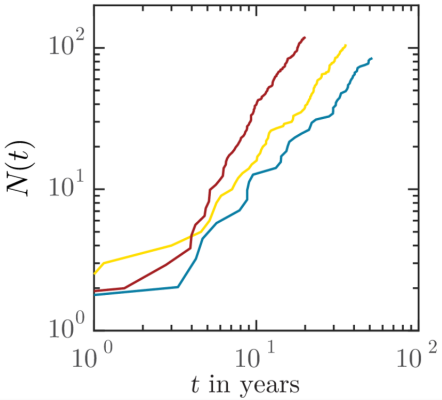
Stratifying Impact

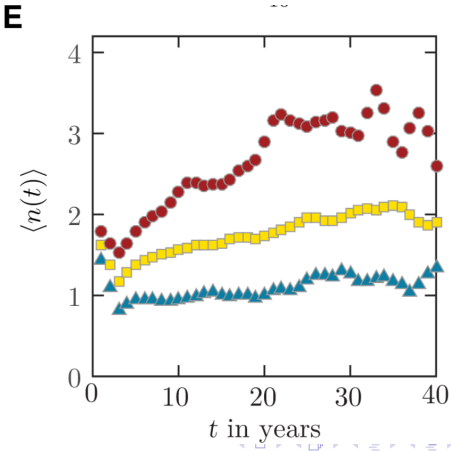
- High maximum impact (top 5%): $c_{10}^* \geq 200$
- Medium maximum impact (middle 75%): $20 < c_{10}^* < 200$
- Low maximum impact (bottom 20%): $c_{10}^* < 20$



Prior Work - Findings

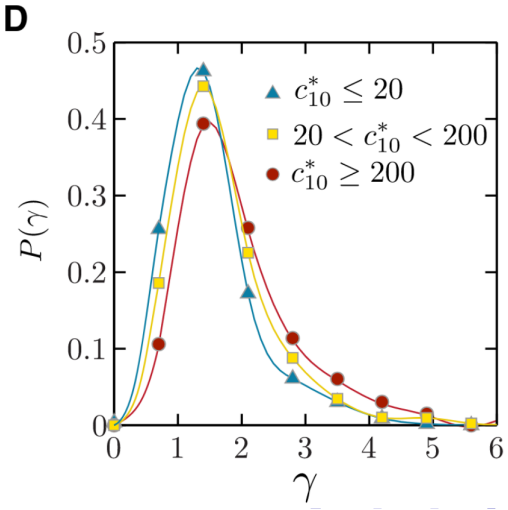
Finding 1: Productivity changes throughout one's career and higher impact scientists gain a greater rate of productivity.



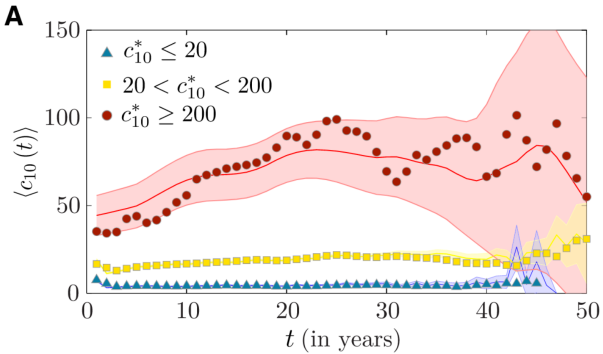


Finding 2: Distribution of Productivity

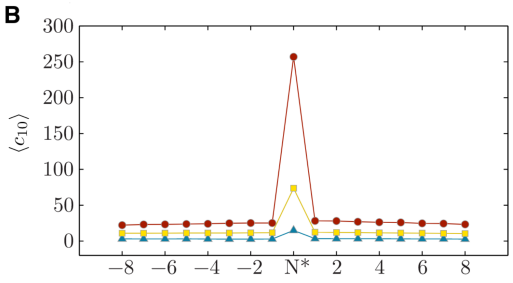
The total number of papers a scientist publishes up to time t after their first publication, $N_i(t)$, asymptotically follows $N_i(t) \sim t_i^{\gamma}$.

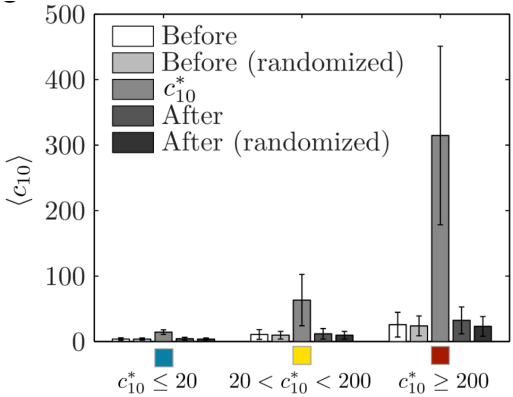


Finding 3: The average impact each year is higher and grows for higher impact scientists.



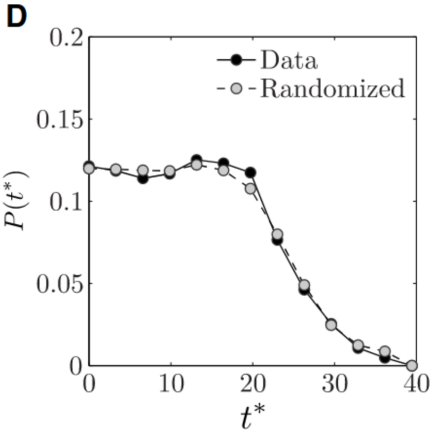
Finding 4: There are no changes in impact leading up to or following a scientist's highest-impact work



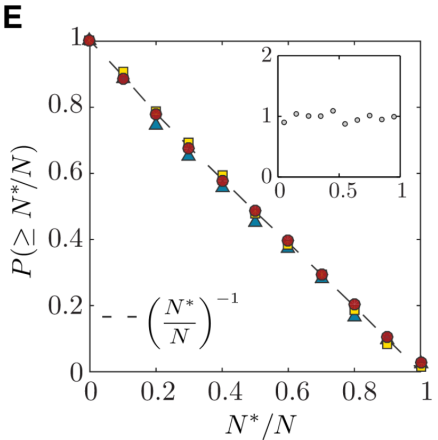


Finding 5: A scientist's highest impact work is typically evenly distributed within the first 20 years of their career.

- The drop after 20 years suggests that it is unlikely that a scientist's most-cited work will come late in her career
- Shuffled c_{10} among all papers published by the same scientist
- Variations are not due to specific impact sequences or other features.



Finding 6: Impact is randomly distributed within a scientist's body of work



Summary of findings so far

1. Productivity changes throughout one's career and higher impact scientists gain a greater rate of productivity.
2. Distribution of Productivity
3. The average impact each year is higher and grows for higher impact scientists.

Summary of findings so far

1. There are no changes in impact leading up to or following a scientist's highest-impact work.
2. A scientist's highest impact work is typically evenly distributed within the first 20 years of their career.
3. Impact is randomly distributed within a scientist's body of work.

What is the role of a researcher's own ability, if any, in scientific excellence?

Random Impact Model (R-Model)

- We assume that each scientist publishes a sequences of papers whose impact is randomly chosen from the same impact distribution $P(c_{10})$.
- Consequently, the only difference between two scientists is their overall productivity N.

This does a great job reproducing the impact distribution $P(N^*/N*)$.

Downsides to the R-Model

1. *Productivity alone begets success:* If each paper's impact is randomly drawn from the same $P(c_{10})$, a productive scientist will more likely score a high c_{10}^*.
2. *Divergent Impact:* Fails to capture the notion that the higher the average impact of a scientist's publications without the most-cited publication, the higher the impact of the most-cited paper.

We need to explore more closely the relationship between **chance**, **productivity**, and **talent**.

Q-Model

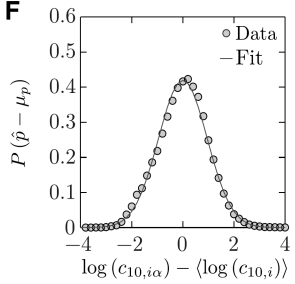
The authors hypothesized that some parameter Q_i individual to each scientist modulates impact.

$$c_{10,i\alpha} = Q_i p_\alpha$$

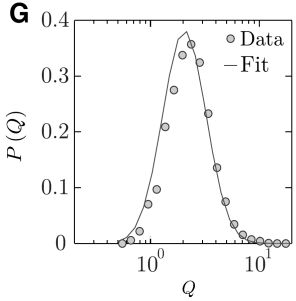
Approximating with Maximum-Likelihood

Approximate the joint probability of $P(\hat{p}, \hat{Q}, \hat{N})$[1] via a trivariate normal distribution.

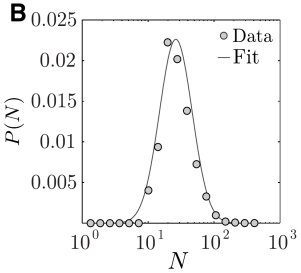
F



G



B



[1] $\hat{x} = \log x$

Q-Model Paramters

$$\mu = (\mu_p, \mu_Q, \mu_N) = (0.92, 0.93, 3.34)$$

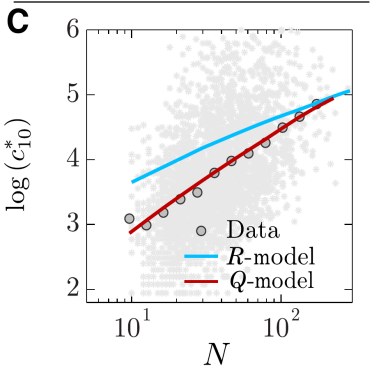
$$\Sigma = \begin{pmatrix} \sigma_p^2 & \sigma_{p,Q} & \sigma_{p,N} \\ \sigma_{p,Q} & \sigma_Q^2 & \sigma_{Q,N} \\ \sigma_{p,N} & \sigma_{Q,N} & \sigma_N^2 \end{pmatrix} = \begin{pmatrix} 0.93 & 0.00 & 0.00 \\ 0.00 & 0.21 & 0.09 \\ 0.00 & 0.09 & 0.33 \end{pmatrix}$$

Key points:

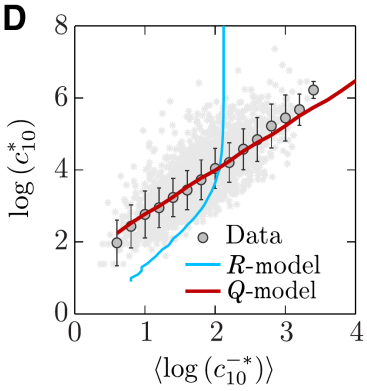
- No relationship betwen paper's potential impact and a scientist's productivity or hidden parameter.
- Slight positive relationship between a scientist's productivity and their hidden factor.

Corrects the downsides of the *R*-model

Productivity begets success



Divergent Impact

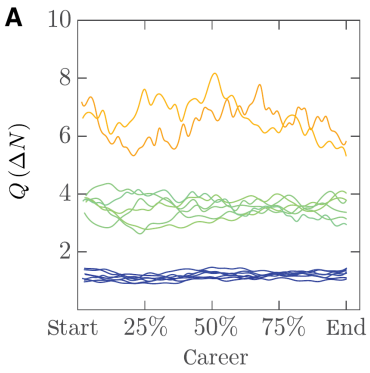


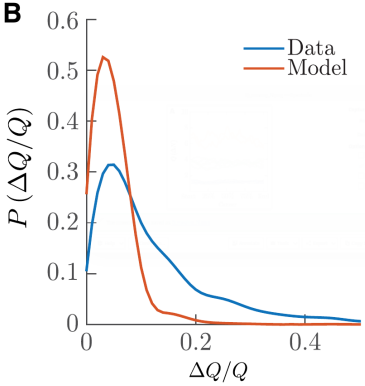
What does this allow us to do?

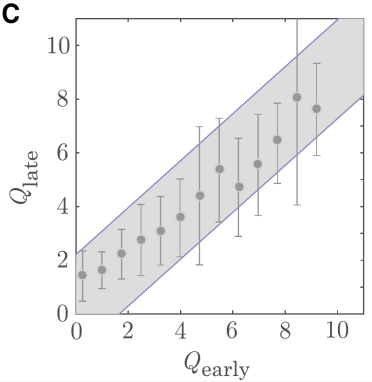
1. Generate synthetic sequences of publications given a scientist's productivity and Q-factor.
2. Estimate a scientist's Q-factor.

$$Q_i = e^{\langle \log c_{10,i} \rangle - \mu p}$$

Stability of Q







Extension to Research

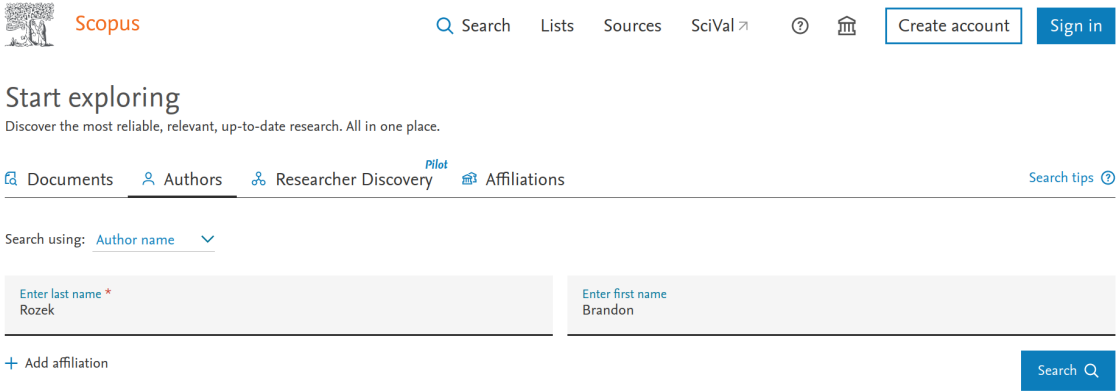
Impact of self-citations

Depending on the field, a project may span multiple publications. Scientists are likely to cite themselves within their papers adding to the citation counts.

How does the usage of self-citations impact these findings?

New Dataset: Scopus

Citation database hosted by Elsevier launched in 2004 and covers 34,346 peer-reviewed journals going back to 1970.



Data Acquisition Process

Retrieved citation data of all the professors in the CS department here at RPI and several of my collaborators.

1. Lookup author
2. Grab 15 year time slices
3. Join data
4. Calculate metrics

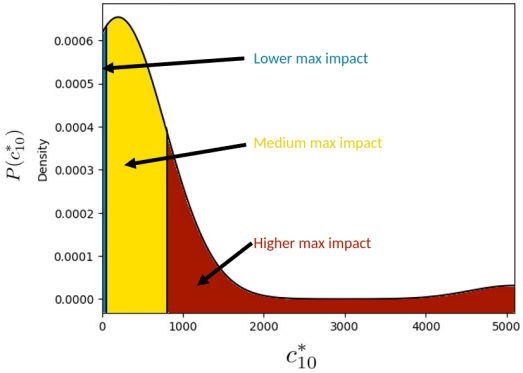
Table: Sample Scopus Data

Title	Year	2020	2021	2022
Paper A	2018	0	0	3
Paper B	2005	10	5	1
Paper C	2020	0	0	0

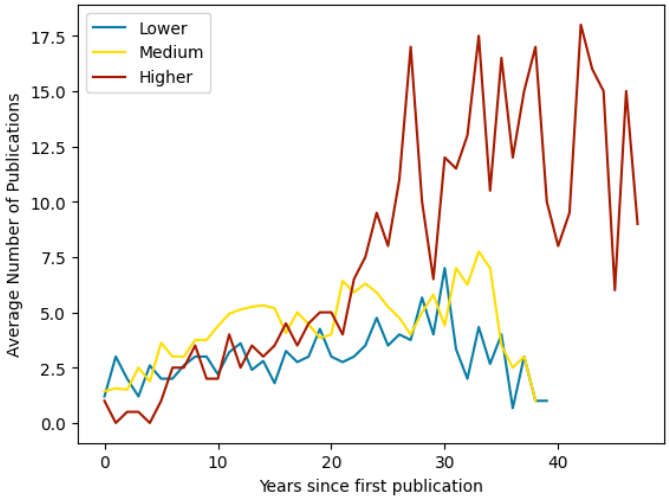
For a total of 23 scientists and 2902 publications.

Redefining Impact Classes

- **Lower Impact**: Less than 60 citations
  - 5 Scientists
- **Medium Impact**: Between 60 and 805 citations
  - 16 Scientists
- **Higher Impact**: More than 805 citations
  - 2 Scientists

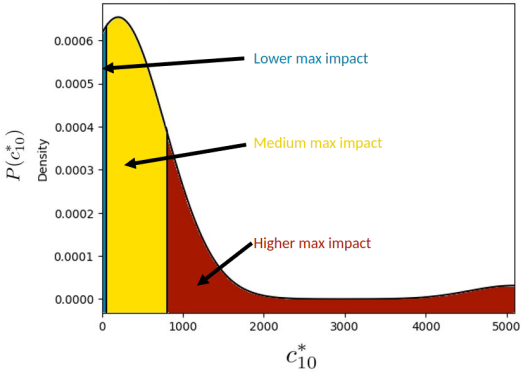


Productivity difference shown in this dataset as well

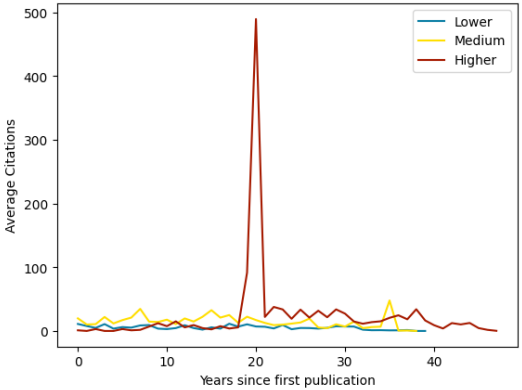


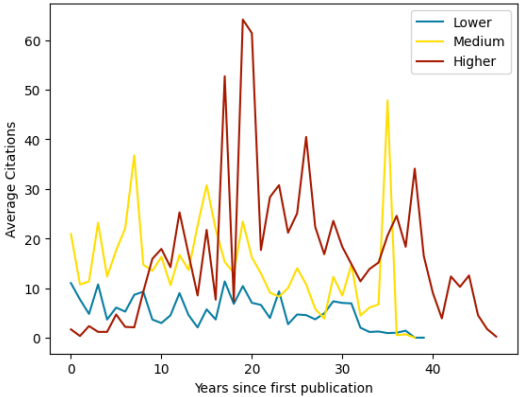
Impact Classes - No Self-Citations

- **Lower Impact**: Less than ~~60~~ 50 citations
  - 5 Scientists
- **Medium Impact**: Between 50 and 794 citations
  - 16 Scientists
- **Higher Impact**: More than ~~805~~ 794 citations
  - 2 Scientists

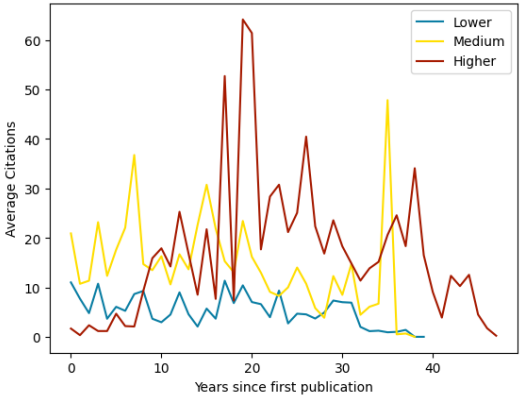


Average Yearly Citation Count by Impact Class

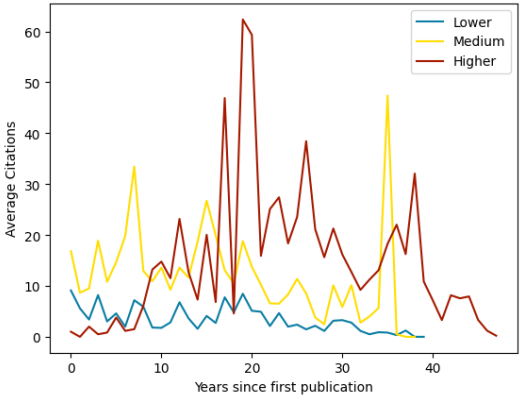




Comparison without self-citations



With self-citation



Without self-citations

Future Work

1. Run statistical tests to see if there's a significant difference between the self-citation allowed versus denied datasets.
2. Replicate random impact hypothesis.

Questions?