Quantifying Long-Term Scientific Impact

Ian Gross
Rensselaer Polytechnic Institute
Frontiers of Network Science - Fall 2017, Assignment 2

grossi2@rpi.edu

November 16, 2017
Overview

1. Background
2. Goal and Purpose of Work
3. Characterizing Citation Dynamics
4. Fundamental Mechanisms of Paper Citation History
   - Equations
5. Validity of Method
6. Evaluating Long-Term Impact
7. Quantifying Changes in a Journal’s Long-Term Impact
8. Prediction of Publication’s Future Citations
   - Uniqueness of Solution
9. Conclusions
Background

- Citation-based metrics is the most frequently used measure of scientific impact
  - Hirsch Index, g-index, impact factors, eigenfactors, and others
- However, our ability to find lasting impact of citation patterns has limitations
  - Historical significance of journals have little impact on internal paper citations
  - Favors older papers and established investigations
  - Hard to predict early citation pattern of well performing papers
  - Incompatible publication, citation, and/or acknowledgement standards
- Past research focus: power law, fat-tail distribution, rescaling of distributions by discipline-dependent variables
- Suggests lack of order and lack of predictability in citation patterns spanning multiple research fields
Goal and Purpose of Work

- Is there long-term predictability in citation papers?
- Derivation of citation dynamics of individual papers
- Collapse citation histories of papers from different journals and disciplines into a single curve
- Uncover basic mechanisms that govern scientific impact and offer potential policy implications
(A) Yearly citations $c_i(t)$ for 200 papers
(B) Average citations acquired 2 years after publication ($c^2$) for papers with same long-term impact ($c^{30}$)
(C) Distribution of aging when cited (log-normal survival)
Fundamental Mechanisms of Paper Citation History

Preferential Attachment
Highly cited papers are more visible and more likely to be cited again.

Aging
New ideas are integrated in subsequent work, so a paper's novelty will fade eventually (log-normal).

\[ P_i(t) = \frac{1}{\sqrt{2\pi \sigma_i t}} \exp\left[-\frac{(\ln t - \mu_i)^2}{2\sigma_i^2}\right] \]

t is time, \( \mu \) is immediacy (time to reach citation peak), \( \sigma \) is longevity (decay rate).

Fitness (\( \eta_i \))
Captures inherent differences between papers, accounting for perceived novelty and importance of discovery (community response).
Probability that paper $i$ is cited at time $t$ after publication (EQ 2):

$$\Pi_i(t) \sim \eta_i c_i^t P_i(t)$$

Predict cumulative number of citations acquired by paper $i$ at time $t$ after publication (EQ 3):

$$c_i^t = m \left[ e^{\frac{\beta \eta_i}{A} \Phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1 \right] \equiv m \left[ e^{\lambda_i \Phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1 \right]$$

Definitions:

- Cumulative Normal Distribution (EQ4): $\Phi(x) \equiv (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy$
- $m =$ avg number of references for each new paper (set to 30)
- $\beta =$ growth rate of total number of publications
- $A =$ Normalization constant
EQ3 represents a minimal citation model, capturing all known quantifiable mechanisms that affect citation histories

Utilizes the three parameters:
- Relative Fitness: $\lambda_i \equiv \frac{n_i \beta}{A}$
- Immediacy: $\mu$
- Rate of Decay: $\sigma$

Using rescaled variables: $\tilde{t} \equiv \frac{(\ln t - \mu_i)}{\sigma_i}$ and $\tilde{c} \equiv \frac{\ln(1+c_i^t)}{\lambda_i}$, we obtain our main result (EQ 5):

$\tilde{c} = \Phi(\tilde{t})$

Prediction: Each paper’s citation history should follow the same universal curve $\Phi(\tilde{t})$, if rescaled with the paper-specific parameters ($\lambda_i$, $\mu_i$, and $\sigma_i$)

Using, paper citation history and EQ 3, we can obtain the best-fitted parameters for paper $i$ (Figures 1D + 1E)

- Figure 1D is yearly citations, Figure 1E is cumulative citations
Validity of Method

- Tested model validly by re-scaling papers between 1950-1980 and found they scaled into EQ 5 (Figures 1F + 1G)
  - 1F is a data collapse, while 1G shows changes in citation history according to changes in Eq3 parameters
- The model Eqs. 3 to 5 predict several measures of impact:
  - **Ultimate Impact** \( (c^\infty) \): Predicts total number of citations acquired by paper during its lifetime. Dependent on relative fitness.
    - \( c^\infty_i = m(e^{\lambda_i} - 1) \)
  - **Impact Time** \( (T^*_i) \): Characteristic time of paper to collect the bulk of its citations. Dependent on immediacy.
    - \( T^*_i \approx \exp(\mu_i) \)
Evaluating Long-Term Impact

Three selected journals of widely varying IFs

- *Proceeding of the National Academy of Sciences USA* (PNAS) (10.48)
- *Cell* (33.62)
Evaluating Long-Term Impact - Findings

- They followed different paths, but converged around year 20 for the cumulative number of citations (2B).
- Having a similar fitness value, they had the same ultimate impact: $c^\infty = 51.5$.
- In the long run, difference in citation count vanishes with time.
- Papers with same citation count at year 2 tend to deviate.
- Fitness and Ultimate Impact offer a journal independent measure of a publication's long-term impact.
The model (Eqs. 3 to 5) helps to connect to IF, the traditional measure of scientific impact, to the journal’s Λ, M, and Σ parameters (the analogs of λ, μ, and σ).

Interesting situation: IF of *Cell* = 38.7 and IF of *NEJM* = 28.7 in 1988. After a decade, *NEJM* increased to 50, while *Cell* decreased to 30.

- Caused by changes in impact time $T^* = \exp(M)$ (Figure 3C)

**Conclusion**: Although both journals show a shift to higher-fitness (Figure 3G), *Cell* shifts to higher-μ papers while *NEJM* remains unchanged.
Quantifying Changes in a Journal’s Long-Term Impact

A 60 50 40 30 20 10 0
Impact Factor

B 800 600 400 200 0
C

C 6 5 4 3 2 1 0

T [year]

D

E 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0

Σ

F 7.5 7.0 6.5 6.0 5.5 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0

M

G 0.5 0.4 0.3 0.2 0.1 0.0

P(λ)

H

π 10 8 6 4 2 0

P(μ)
Prediction of Publications Future Citations

- Framework borrowed from weather predictions and data mining
- Used paper $i$’s citation history up to year $T_{Train}$ after publication (training period) to estimate $\lambda$, $\mu$, and $\sigma$
- Then used Eq.3 to predict citations $c_i^T$ and Eq.6 to determine its ultimate impact $c_i^\infty$
- Assigned envelope to each paper to explicitly quantify uncertainty
- To quantify predictive accuracy, they measured the fraction of papers that fall within the envelope for all PR papers published in 1960s
  - Result: The model correctly approximated the citation range for 93.5% of papers 25 years into the future
Identified several models used in the past to fit citation histories

Fit the predictions of these models to PR papers and used Kolmogorov-Sirnov (KS) test to evaluate their goodness of fit

- Captures max deviation between the fitted and empirical data

Lowest KS distribution: Eq3, shows best fit (Fig. 4D)

Eq3 was able to predict more accurate results than past models (Fig. 4E & 4F)

Log-normal model correctly captures citation of small impact papers, yet fails for medium to high impact papers

- Allows us to predict citation threshold for when preferential attachment becomes relevant ($\lambda < 0.25$ and $c^\infty < 8.5$)
Conclusions

- Limitations:
  - Cannot account for exogenous second acts (ex: a new discovery)
  - Delayed Impact (ex: Erdős and Rényi’s work)

- Proposed model can impact the current citation-based metrics integrated into reward procedures (grants, awards, bonuses, etc.)

- IF and short-term citations lack predictive power

- $c^\infty$ offers a journal-independent assessment of a paper’s long term impact
  - Captures total number of citations a paper will ever acquire or the discovery’s ultimate impact

- Additional variables with data mining could be used to enhance predictive power

- Helps the community understand the factors that govern a research community’s response to a discovery
Dashun Wang, Chaoming Song, Albert-László Barabási
Quantifying Long-Term Scientific Impact
Science 04 Oct 2013
Vol. 342, Issue 6154, pp. 127-132
Questions?